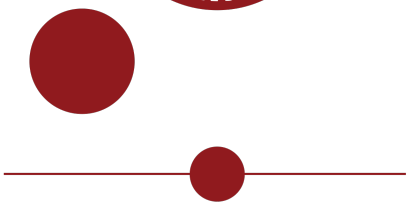
RAHUL ARALIKATTE

TOWARDS BETTER NEURAL COREFERENCE RESOLUTION

This thesis has been submitted to the Ph.D. School of The Faculty of Science, University of Copenhagen.

TOWARDS BETTER NEURAL COREFERENCE RESOLUTION

RAHUL ARALIKATTE



KØBENHAVNS
UNIVERSITET

Ph.D. Thesis
September 2021

SUPERVISOR:
Anders Søgaard

ASSESSMENT COMMITTEE:
Desmond Elliott, University of Copenhagen
Ivan Vulić, University of Cambridge
Adina Williams, Facebook AI Research

AFFILIATION:
Department of Computer Science
Faculty of Science
University of Copenhagen

THESIS SUBMITTED:
September 2021

ABSTRACT

Though ambiguity is inherent in language, humans are adept at resolving them from explicit context or implicit knowledge. Computer algorithms, on the other hand, find it hard to resolve ambiguous constructions like anaphora. This has changed with the introduction of Deep Neural Networks to NLP. These networks, which contain millions of parameters trained on billions of data points, can now solve the aforementioned problems to a large extent, especially in high-resource languages like English. However, their performance is bound by the availability of labeled data. Increasing their performance requires further expansion of model capacity or curation of larger datasets which are prohibitively expensive and give diminishing returns. Therefore, we need to find better ways of improving these networks without relying on labeled data alone.

In this work, we propose three new methods to improve coreference resolution: (1) Augmenting external knowledge: knowledge bases contain enormous amounts of real-world knowledge which can be used to resolve ambiguities that arise from grounding assumptions. We introduce a reinforcement learning-based approach that improves performance by verifying model decisions against external knowledge bases and rewarding them based on their validity (Chapter 2), (2) Remodelling of tasks: performance of some tasks can be improved if they are recast into a different form that is more suitable for learning. Since the availability of training data varies drastically across tasks, we remodel a low-resource task to take the form of a high-resource task, and use models pre-trained for the latter and finetune them to get significant improvements on the former (Chapter 3), and (3) Encouraging coherence in MTL: in standard multitask learning setups, strongly correlated tasks result in better overall performance. Taking this a step further, we build simple meaning representations from the outputs of the model to explicitly quantify the coherence between them and use this coherence value as a reward to further finetune the models (Chapter 4). We thoroughly experiment with and analyze these methods and report performance improvements across the board.

Finally, in the last part of this work, we introduce two general methods which can be used to improve a variety of NLP tasks: (1) Focus Attention: we introduce a new kind of attention mechanism that enhances the faithfulness of transformer-based seq2seq models by biasing the decoder to generate text which is thematically consistent with the input. We show that this mechanism improves the faithfulness of state-of-the-art abstractive summarization systems (Chapter 6), and (2) Robust MAML: we improve the vanilla Model Agnostic Meta-learning algorithm by introducing two new criteria, which either minimizes the maximum risk across tasks or constrains the risk of each task to be below a threshold. We evaluate these criteria on

POS-tagging and question-answering to find that they work exceptionally well for out-of-distribution transfer, especially in zero- and few-shot settings (Chapter 7).

ABSTRAKT

Mennesker er kompetente til at navigere i sproglige tvetydigheder, hvor betydningen skal udledes af kontekst eller kræver implicit viden. Computeralgoritmer har derimod svært ved at gennemskue meningen bag sproglige virkemidler som anaforer. Dette har ændret sig med introduktionen af *Deep Neural Networks* i NLP. Disse netværk, der indeholder millioner af parametre trænet på milliarder af observationer, kan nu i høj grad løse førnævnte udfordringer især for sprog som engelsk, hvor der findes et stort antal tilgængelige ressourcer. Resultatet er dog dybt afhængigt af tilgængeligheden af observationer for hvilken den sande betydning er kendt. At opnå bedre resultater kræver yderligere udvidelse af modellernes kapacitet eller indsamling af datasæt i en størrelsesorden, hvor de bliver uoverkommelige at organisere og giver faldende afkast. Derfor er vi nødt til at finde bedre måder at forbedre disse netværk uden at være afhængig af data, hvor den sande betydning er kendt.

Denne afhandling foreslår tre nye metoder til at forbedre *coreference resolution*: (1) Forøgelse af ekstern viden: vidensbaser indeholder enorme mængder af viden fra den virkelige verden, der kan bruges til at løse uklarhed, der opstår på grund af grundantagelser. Vi introducerer en tilgang baseret på *reinforcement learning*, der fører til bedre resultater ved at holde modelbeslutninger op mod eksterne vidensbaser og belønne dem baseret på deres gyldighed (Kapitel 2). (2) Omformning af opgaver: resultatet af nogle opgaver kan forbedres, hvis opgaverne gives en anden form, der er bedre egnet til læringsbaserede løsninger. Da tilgængeligheden af træningsdata varierer drastisk på tværs af opgaver, omformer vi en opgave ellers kendetegnet ved et lavt antal ressourcer til en kendetegnet ved et højt antal ressourcer. Vi gør brug af præ-trænede modeller for sidstnævnte og fintuner dem for at opnå en signifikant forbedring af førnævnte (Kapitel 3). (3) Tilskyndelse af sammenhæng i MTL: i standardtilgange til *multitask learning* fører stærkt korrelerede opgaver til et bedre samlet resultat. Ved at tage dette et skridt videre kan vi bygge enkle repræsentationer for betydning ud fra modellens output. Ud fra disse kan sammenhængen mellem dem kvantificeres og denne værdi for sammenhæng kan bruges som belønning til videre fintuning af modellerne (Kapitel 4).

I den sidste del af denne afhandling introducerer vi generelle metoder, der kan bruges til at forbedre en række af opgaver i NLP: (1) *Focus Attention*: vi introducerer en ny form for mekanisme til opmærksomhed, der øger troværdigheden af *transformer*-baserede seq2seq modeller ved at bias*decoder*. Målet er at generere tekst, der er tematisk i overensstemmelse med input. Vi viser, at denne mekanisme

forbedrer troværdigheden i state-of-the-art abstrakte opsummeringssystemer (Kapitel ref ch: 05). (2) Robust MAML: vi forbedrer den grundlæggende *Model Agnostic Meta-learning* algoritme ved at introducere to nye kriterier, som enten minimerer den maksimale risiko på tværs af opgaver eller begrænser risikoen i hver opgave til at være under en tærskelværdi. Vi evaluerer disse kriterier på POS-tagging og *question-answering* for at finde ud af, at de overføres exceptionelt godt til uden for fordelingen, især i (zero-) og *few-shot* sammenhænge (Kapitel 7).

PUBLICATIONS

This is an article-based dissertation. The following list of peer-reviewed publications are included in this thesis.

Aralikatte, Rahul, Mostafa Abdou, Heather C Lent, Daniel Hershcovich, and Anders Søgaard (2021a). "Joint Semantic Analysis with Document-Level Cross-Task Coherence Rewards." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.14, pp. 12516–12525. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17484.

Aralikatte, Rahul, Matthew Lamm, Daniel Hardt, and Anders Søgaard (Apr. 2021b). "Ellipsis Resolution as Question Answering: An Evaluation." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 810–817. URL: https://aclanthology.org/2021.eacl-main.68.

Aralikatte, Rahul, Heather Lent, Ana Valeria Gonzalez, Daniel Herschcovich, Chen Qiu, Anders Sandholm, Michael Ringaard, and Anders Søgaard (Nov. 2019). "Rewarding Coreference Resolvers for Being Consistent with World Knowledge." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 1229–1235. DOI: 10.18653/v1/D19-1118. URL: https://aclanthology.org/D19-1118.

Aralikatte, Rahul, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald (Aug. 2021c). "Focus Attention: Promoting Faithfulness and Diversity in Summarization." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 6078–6095. URL: https://aclanthology.org/2021.acl-long.474.

Aralikatte, Rahul and Anders Søgaard (May 2020). "Model-based Annotation of Coreference." English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 74–79. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.9.

Ponti, Edoardo Maria, Rahul Aralikatte, Disha Shrivastava, Siva Reddy, and Anders Søgaard (Aug. 2021). "Minimax and Neyman–Pearson Meta-Learning for Outlier Languages." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 1245–1260. URL: https://aclanthology.org/2021.findings-acl.106.

Below is a list of publications and manuscripts co-authored by me during the course of my Ph.D., but **not** included as part of this thesis.

Abdou, Mostafa, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein, and Anders Søgaard (Nov. 2019). "X-WikiRE: A Large, Multilingual Resource for Relation Extraction as Machine Comprehension." In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, pp. 265–274. DOI: 10.18653/v1/D19-6130. URL: https://aclanthology.org/D19-6130.

Aralikatte, Rahul, Miryam de Lhoneux, Anoop Kunchukuttan, and Anders Søgaard (Aug. 2021a). "Itihasa: A large-scale corpus for Sanskrit to English translation." In: *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*. Online: Association for Computational Linguistics, pp. 191–197. URL: https://aclanthology.org/2021.wat-1.22.

Aralikatte, Rahul, Héctor Ricardo Murrieta Bello, Daniel Hershcovich, Marcel Bollmann, and Anders Søgaard (Aug. 2021b). "How far can we get with one GPU in 100 hours? CoAStaL at MultiIndicMT Shared Task." In: *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*. Online: Association for Computational Linguistics, pp. 205–211. URL: https://aclanthology.org/2021.wat-1.24.

Bollmann, Marcel, Rahul Aralikatte, Héctor Murrieta Bello, Daniel Hershcovich, Miryam de Lhoneux, and Anders Søgaard (June 2021). "Moses and the Character-Based Random Babbling Baseline: CoAStaL at AmericasNLP 2021 Shared Task." In: *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Online: Association for Computational Linguistics, pp. 248–254. DOI: 10.18653/v1/2021.americasnlp-1.28. URL: https://aclanthology.org/2021.americasnlp-1.28.

Nikolaus, Mitja, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott (Nov. 2019). "Compositional Generalization in Image Captioning." In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 87–98. DOI: 10.18653/v1/K19-1009. URL: https://aclanthology.org/K19-1009.

Samprāpte sannihite kāle nahi nahi rakshati dukrnkarane.
(When death nears, no knowledge of grammar can save you.)[1]
— Adi Shankara: 810 CE

ACKNOWLEDGMENTS

No man is an island. Ph.D. is an endeavor that requires a support system comparable to that of an Olympian. There are so many people who have, knowingly or otherwise, helped me get to this point. I want to thank every one of those people starting with Anders Søgaard. I could not have asked for a better supervisor, especially when the world ground to a halt in 2020. He gave me the freedom and encouragement to work on a whole range of diverse topics. I also thank the entire CoAStaL crew, who are a bunch of amazing, kindhearted, and brilliant people. My Ph.D. would not be as memorable as it was without them.

I have had the fortune of working with a long list of wonderful collaborators both from academia and the industry. All of them have had a significant role in molding me into a (hopefully) well-rounded researcher. I especially thank Mostafa Abdou, Daniel Hershcovich, Heather Lent, and Edoardo Ponti for doing the heavy lifting during our time working together. Also, I owe a debt of gratitude to Shashi Narayan, who worked so hard to make my Google internship fruitful.

I would also like to thank my parents and grandparents who have been nothing but supportive and have sacrificed so much for me. Finally, I thank Vidhyashree Murthy, who has stood by me through thick and thin and has been a constant companion in all my adventures since we were younglings.

1 Helps me remember that there are more important things than linguistics and NLP.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

BACKGROUND

1

INTRODUCTION

Human language is inherently vague and highly ambiguous if viewed without context. Over centuries, we have developed mechanisms like *anaphora*[1] to communicate efficiently by eliminating syntactic redundancies while maintaining semantic clarity. In particular, coreference allows us to use shorter words (pro-forms) instead of repeating longer expressions. For example,

John and Mary went to the market. He bought some shoes while she browsed around. They went home soon after.

Here, 'John' in the first sentence is referred to as 'he' in the second sentence. The same holds for 'Mary' and 'she'. In the third sentence, 'they' refers to both 'John' and 'Mary'. Formally, we say an expression refers to another, or two expressions co-refer each other if they represent the same *referent*. Referents are usually entities like people or things. The first expression where the referent is described without using pro-forms is known as an *antecedent*[2]. All other references to the antecedent are called *mentions*. In the above example, 'John' is an antecedent, and 'he' and 'they' are mentions.

Resolving coreference can require contextual information available explicitly or implicit knowledge of the world. Let us look at each case briefly to understand why the latter is much harder for computers than the former.

EXPLICIT CONTEXT In this case, all information required to resolve the coreference is explicitly made available by the speaker. For example, in the example given above, we can easily resolve 'he' as referring to 'John' by using the information present in the first sentence.[3] We can also write simple rules to make a computer identify such references. However, this is not true for cases requiring real-world knowledge.

IMPLICIT KNOWLEDGE Our brain stores huge amounts of knowledge related to the world around us, which is also consistent with that of other people. This enables us to ground our communications easily using pro-forms without having to provide detailed context. For example, if we ask someone, 'Is it hot outside?', we assume that the listener understands that 'it' refers to a space external to the current environment. The assumption is made since the speaker is confident that they share a mental model of the world with the listener.[4] An-

1 Here, anaphora is used in a broad sense, to include both anaphora and cataphora.

2 If the referent occurs after the mentions, then it is known as a *postcedent*.

3 We require certain world-knowledge to identify that 'John' is a better-suited antecedent than 'Mary' because 'he' is a masculine pronoun. However, let us choose to ignore it in this case.

4 These kind of references are known as *indirect* or *associative* anaphora.

swering this question requires additional assumptions such as what combinations of air temperature and humidity constitutes 'hot', etc. On the other hand, communicating these assumptions with computers are non-trivial (Staggers and Norcio, 1993). An easier workaround is to convert implicit knowledge into programming literals and store them in memory. This allows us to check their values before making decisions. For example,

```
if temp > 25 and humidity > 0.5:
    print("It is hot!")
else:
    print("Not so much.")
```

In general, using such knowledge to disambiguate pro-forms is highly challenging. Therefore identifying and resolving anaphoric ambiguities is considered to be one of the foundational goals of language research. The rest of the chapter gives an overview of important previous works on coreference resolution, stresses the importance of coreference in today's NLP landscape, and enumerates research contributions made in this thesis.

1.1 PREVIOUS WORK

Classically, automatic coreference resolution relied heavily on linguistic theories (Elango, 2005). Soon, statistical methods were developed to find language patterns, which became features for machine learning algorithms. We shall now briefly look at essential works from both categories.

1.1.1 *Linguistic Approaches*

HOBB'S ALGORITHM Being one of the earliest approaches, it mainly used syntactic parse trees of sentences to resolve pronouns (Hobbs, 1986). Initially, the algorithm tries to find antecedents within the current sentence of interest using a breadth-first search of the parse tree. Explicit rules are written to accommodate nuances like *contra-indexing*. Finally, parse trees of previous sentences are searched for antecedents in reverse chronological order using the same method. Overall, this method prefers antecedents that are closer to the mentions.

CENTRING BASED METHODS These methods get their name from being built on top of *Centring Theory* (Grosz et al., 1995). Centering theory is a general framework for tracking the focal points or *centers* of utterances. The Brennan-Friedman-Pollard (BFC, Brennan et al., 1987) method works with the assumption that pro-forms help readers focus their attention. This idea is used to find antecedents using forward and backward *centers* of sentences. They construct all possible backward-forward pairs and filter them based on certain constraints. Then, they classify each pair based on hand-written rules and select

the highest-scoring pairs, the backward centers of which become the antecedents of the forward centers.

The Left-Right-Centering (Tetreault, 1999) algorithm refines BFC by eliminating the need to generate all backward-forward pairs, which is computationally expensive. It supports an incremental resolution that is inspired by Hobb's algorithm. It starts by searching for the antecedent in the same sentence by looking at candidate forward-centers that meet particular features and binding constraints. If not found, the search is expanded iteratively to previous sentences. Many other pronoun resolution methods have centering theory at their core, such as (Kong et al., 2009; Uryupina, 2006).

Approaches like Strube, 1998 do not use centering in the traditional sense. However, they use major ideas from it like entity hierarchies. They maintain lists of entities already seen, and when a pronoun is encountered, the entities are ranked using standard constraints such as binding constraints, etc. This is more natural and how humans interpret pro-forms.

METHODS FOR BRIDGING REFERENCES Associative anaphora, whose antecedents are not directly mentioned in the discourse require bridging references. As elaborated previously, these kinds of references require background knowledge. Early works like (Rahman and Ng, 2011a) used knowledge bases like YAGO (Suchanek et al., 2007) or WordNet (Miller, 1995) to determine whether any connection exists between candidate antecedents and the referring expression. Systems like (Bunescu, 2003; Markert et al., 2003; Poesio et al., 2004) used search engines to retrieve documents containing the referring expression and up-weighted those antecedents which appeared in them.

1.1.2 *ML-based Approaches*

Most classic ML approaches involved manually extracting features from discourse and using them to train Naive Bayes (Ge et al., 1998), Decision Trees (Soon et al., 2001), Conditional Random Fields (McCallum and Wellner, 2005), etc. These works introduce many heuristics that are shown to work well on test sets like MUC.[5] The most important features, which are common among in these works are:

- Distance between the mention and a potential antecedent
- Syntactic structure, which enables resolving standard constraints
- Agreement constraints, which make sure both mentions and antecedents have plausible gender forms, grammatical number, animacy, etc.
- Semantic class agreement, which verifies the compatibility between mention and antecedent's semantic classes, using an external resource like WordNet

5 https://catalog.ldc.upenn.edu/LDC2003T13

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q1: What are **the candidates** running for?
A1: Governor
Q2: Where?
A2: Virginia
Q3: Who is the democratic candidate?
A3: Terry McAuliffe
Q4: Who is **his** opponent?
A4: Ken Cuccinelli
Q5: What party does **he** belong to?
A5: Republican
Q6: Which of **them** is winning?
A6: Terry McAuliffe

Figure 1: An example conversation from the CoQA dataset.

- Syntactic similarity, which measures if two expressions are different forms of the same word(s)

CLUSTERING In works like (Cardie and Wagstaff, 1999), clustering algorithms are applied to feature representations of noun phrases. When a noun phrase is added to a cluster or when two clusters are merged, a consistency check is performed across all members of the cluster(s). This causes large compute overhead, especially for long documents. (Wagstaff and Cardie, 2002) handle this by pre-computing specific constraints between noun phrases. For example, the 'cannot-link' constraint if the genders do not match or the 'must-link' constraint if they have significant lexical overlap. These constraints are not perfect and can bias the clusters in undesirable ways.

1.2 COREFERENCE AND MODERN NLP

With the advent of LSTMs (Hochreiter and Schmidhuber, 1997b), and the more recent Transformers (Vaswani et al., 2017), coreference resolution research has now stagnated to learning better span representations (Joshi et al., 2019a,b; Lee et al., 2018a). However, recent works show that incorporating external knowledge often improves these models significantly (Zhang et al., 2019a). This line of research is also promising because of its direct application to other domains like dialog and QA, as elaborated next.

COREFERENCE AS A STEPPING STONE Availability of large datasets Common Crawl[6] and C4 (Raffel et al., 2019b) and compute have resulted in the development of pre-trained language models (Devlin et al., 2018; Lewis et al., 2019; Liu et al., 2019) which mimic human language with great fluency. These models can be applied to a variety of tasks to get exceptional performance. They even perform on-par with or beyond human-level on classic NLP tasks like tagging and classification. Therefore, language research is now slowly moving towards more challenging tasks like dialog and question-answering. We see that current models fail to achieve decent results on these harder tasks, especially if they are set in an open domain.

6 https://commoncrawl.org/

Improving coreference resolvers will have a direct impact on the performance of these tasks. For example, consider a sample conversation from the CoQA (Reddy et al., 2019) dataset as shown in Fig. 1. Here, we find many mentions referring to different entities. Failing to link them correctly will result in wrong answers. This is a harder problem because the focus of the conversation keeps shifting with time, at a faster pace than in other types of discourse. The entity focus changes at Q4, Q5, and Q6. Consequently, 'his' and 'he' refers to 'Terry' and 'Ken' respectively, at different points in the conversation.

1.3 RESEARCH CONTRIBUTIONS

In this thesis, we mainly look at different methods of improving coreference resolution (Part ii). Rather than taking the more often-tread path of learning *better* span representations using bigger models, we experiment with three major ideas which build on existing state-of-the-art.

EXTERNAL KNOWLEDGE It is well-known that coreference resolution helps improve relation extraction. In Chapter 2, we show that the opposite is also true, i.e. we can improve coreference resolution by extracting subject-relation-object (SRO) triples from coreference resolved text and verifying them against an external knowledge base.

We reward pre-trained coreference resolvers if their resolutions result in SRO triples that are consistent with world knowledge. We generalize the consistency signal beyond a knowledge base's coverage by training classifiers that predict the probability of an SRO triple being true. These classifiers are inspired by Universal Schema models, which were originally used for relation prediction (Verga and McCallum, 2016b). The resolvers are finetuned using a policy gradient algorithm with the classifier probability as the reward.

TASK REDESIGN We explore if task performance can be improved by changing its structure. In general, anaphoric and elliptic constructions can easily be converted into questions. Their resolutions are often the answer to those questions. Therefore, in Chapter 3, we recast ellipsis and coreference resolution as question-answering. We show that this recasting indeed improves the state-of-the-art for both verb-phrase and sluice ellipsis by a significant margin. However, coreference improvements turn out to be marginal, leading us to postulate that this method is useful when a task is data deficient.

ENCOURAGING COHERENCE In Chapter 4, we explore how semantic role labels (SRL) and coreference links can be combined to build simple graph structures to represent discourse. We find a way to quantify the coherence of these graphs and use them as reward signals to finetune both the SRL and coreference labelers. This finetuning is done in a semi-supervised manner which does not require any additional labeled data.

Pre-trained coreference and SRL models are used to annotate free text, which is then converted into semantic graphs. The coherence scores of these graphs are used as rewards to finetune both models using policy gradient. Coherence scores are computed by coherence classifiers which are trained on the same data on which the coreference and SRL models are pre-trained.

OTHER TOPICS In Part iii of the thesis, we focus on more general methods that can be applied to most NLP tasks, including coreference resolution. In Chapter 6, we introduce a new kind of attention mechanism called *Focus Attention*. Focus attention conditions transformer-based seq2seq decoders to proactively generate tokens which are thematically similar to the input. We also introduce *Focus Sampling*, a controllable sampling mechanism that enhances the diversity of the generated output while retaining its faithfulness to the input. We apply these techniques to abstractive summarization to find that they make state-of-the-art models hallucinate less and do not hinder their performance when measured on standard metrics based on lexical overlap.

Finally in Chapter 7, we introduce two new flavors of MAML which break vanilla MAML's i.i.d assumption. Instead of minimizing the expected risk across all tasks, the new methods can control task risks in a more fine-grained manner. Minimax MAML minimizes the maximum risk across tasks, and Neyman-Pearson MAML constrains the risk of a task to a maximum threshold. These methods are applied to POS-tagging and QA models, where the training and evaluation data are carefully chosen to have a large distributional gap. We find that the proposed methods significantly outperform vanilla MAML and strong multitask baselines.

Part II

COREFERENCE RESOLUTION

2

REWARDING COREFERENCE RESOLVERS FOR BEING CONSISTENT WITH WORLD KNOWLEDGE

2.1 ABSTRACT

Unresolved coreference is a bottleneck for relation extraction, and high-quality coreference resolvers may produce an output that makes it a lot easier to extract knowledge triples. We show how to improve coreference resolvers by forwarding their input to a relation extraction system and reward the resolvers for producing triples that are found in knowledge bases. Since relation extraction systems can rely on different forms of supervision and be biased in different ways, we obtain the best performance, improving over the state of the art, using multi-task reinforcement learning.

2.2 INTRODUCTION

Coreference annotations are costly and difficult to obtain, since trained annotators with sufficient world knowledge are necessary for reliable annotations. This paper presents a way to *simulate* annotators using reinforcement learning. To motivate our approach, we rely on the following example from Martschat and Strube (2014, underlines added to mark entity mentions):

(1) [Lynyrd Skynyrd]$_1$ was formed in Florida$_2$. Other bands from [the Sunshine State]$_2$ include Fireflight and Marilyn Manson.

Martschat and Strube (2014) cite the association between Florida and the Sunshine State as an example of a common source of name-name recall error for state-of-the-art coreference resolution systems. The challenge is that the two names co-occur relatively infrequently and are unlikely to do so in a moderate-sized, manually annotated training corpus. A state-of-the-art system may be able to infer the relation using distributional information about the phrase the Sunshine State, but is likely to have limited evidence for the decision that it is coreferential with Florida rather than Lynyrd Skynyrd.

While coreference-annotated data is scarce, knowledge bases including factual information (such as that Fireflight is from Florida) are increasingly available. For a human annotator unaware that Florida is sometimes referred to as the Sunshine State, the information that Fireflight is from Florida is sufficient to establish that Florida and the Sunshine State are (with high probability) coreferential. This paper explores a novel architecture for making use of such information from knowledge bases by tying a coreference resolution system to a relation extraction system, enabling us to reward the coreference system for making predictions that lead us to infer facts that are consistent

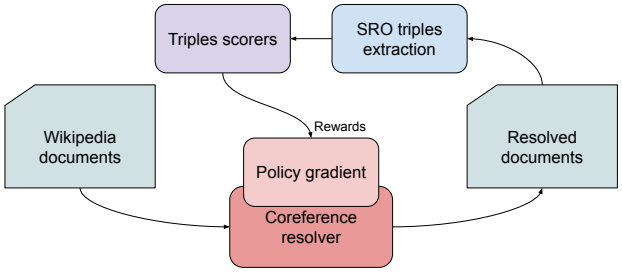


Figure 2: Our strategy for training a coreference resolver using reward from relation extraction.

with such knowledge bases. This potentially provides us with more evidence for resolving coreference such as (1).

We propose a training strategy (Figure 2) in which we pass on the predictions of a neural coreference resolver to an openRE system, matching relations extracted from resolved sentences with a knowledge base. We show how checking the produced relationships for consistency against the knowledge base produces a reward that is, indirectly, a signal about the quality of the coreference resolution. In order to generalize this signal beyond the coverage of the knowledge base, we train a Universal Schema model (Riedel et al., 2013) and use its confidence as our reward function. With this reward function, we do policy-gradient fine-tuning of our coreference resolver, effectively optimizing its predictions' consistency with world knowledge.

CONTRIBUTIONS We demonstrate that training a coreference resolver by reinforcement learning with rewards from a relation extraction system, results in improvements for coreference resolution. Our code is made publicly available at https://github.com/rahular/coref-rl

2.3 CONSISTENCY REWARD FOR COREFERENCE RESOLUTION

In order to reward a coreference resolver for being consistent with world knowledge, we propose a simple training strategy based on relation extraction: (i) Sample a Wikipedia[1] document at random, (ii) Replace mentions with their antecedents using a coreference resolver, (iii) Apply an off-the-shelf openRE system to each rewritten document, (iv) Score relationships that include coreferent mentions using Universal Schema, and (v) Use the score as a reward for training the coreference resolvers.

2.3.1 *Reward functions*

To model consistency with world knowledge, we train different Universal Schema models (Riedel et al., 2013; Verga and McCallum, 2016a), resulting in three reward functions (Figure 3): **RE-KG** (Knowledge Graph Universal Schema) is trained to predict whether two entities

1 https://www.wikipedia.org

	RE-KG	RE-Text	RE-Joint
Document	NA	Bach was a German composer. He is known for instrumental compositions such as the Art of Fugue	Bach was a German composer. He is known for instrumental compositions such as the Art of Fugue
OpenRE	NA	(Bach, born in, Germany) (Bach, occupation, Composer) (He, composition, Art of Fugue)	(Bach, born in, Germany) (Bach, occupation, Composer) (He, composition, Art of Fugue)
Verification	NA	NA	✗ (Bach, born in, Germany) ✓ (Bach, occupation, Composer) ✗ (He, composition, Art of Fugue)
Result	(Bach, born in, Eisenach) (Bach, born on, 21 March 1685) (Bach, occupation, Composer) …	(Bach, born in, Germany) (Bach, occupation, Composer) (He, composition, Art of Fugue)	(Bach, occupation, Composer)

Figure 3: The columns show the different pipelines used to obtain data for training the reward models. The pipeline for: (i) RE-KG directly extracts triples from Wikidata, (ii) RE-Text runs Wikipedia summaries through OpenRE to generate triples, and (iii) RE-Joint adds an additional verification step by checking if the generated triples exist in Wikidata.

are linked in Wikidata[2]; **RE-Text** (Text-based Universal Schema) is trained to predict whether two entities co-occur in Wikipedia; and **RE-Joint** (Joint Universal Schema) is trained to predict whether two entities are linked *and* co-occur. The three rewards focus on different aspects of relationships between entities, giving complimentary views of what entities are related.

Similar to Verga et al. (2016), we parameterize candidate relation phrases with a BiLSTM (Graves and Schmidhuber, 2005), and use pre-trained Wikidata BigGraph embeddings (Lerer et al., 2019) as the entity representations. We apply an MLP with a single hidden layer on the concatenated representations to get the reward value.

2.3.2 *Updating the coreference resolver*

Each resolved document is converted into n SRO triples by an open information retrieval system (Angeli et al., 2015). Each triple t_i is then scored using a reward function to obtain a reward r_i for $i \in \{1, \ldots, n\}$. The final document-level reward is the normalized sum of the individual rewards as shown in Equation 2.1, where R_h is a moving window containing the previous $h = 100$ normalized reward values.

$$R = \frac{\sum_i r_i - mean(R_h)}{stddev(R_h)} \tag{2.1}$$

Since R is not differentiable with respect to the coreference resolver's parameters, we use policy gradient training to update the coreference resolver. We select the best action according to the current policy, using random exploration of the alternative solutions with $p = \frac{1}{10}$.

2 https://www.wikidata.org

2.3.3 *Multi-task reinforcement learning*

Our overall training procedure is presented in Algorithm 1. After training the three aforementioned reward models, we create **RE-Distill** by interpolating their trained weights. Next, we pre-train a coreference resolver using supervised learning, and fine-tune it using each of the three reward functions to get three different coreference policies: **Coref-KG**, **Coref-Text** and **Coref-Joint**, respectively. We then use multi-task reinforcement learning to combine these three policies to get **Coref-Distill**. Our approach is a particular instance of DisTraL (Teh et al., 2017), using policy gradient and model interpolation. Finally, **Coref-Distill** is fine-tuned with rewards from **RE-Distill**.

Algorithm 1 Multi-task Reinforcement Learning

Require: Baseline initialized policies θ_n for $n \in \{1, 2, 3\}$
Require: Reward functions $\texttt{reward}_n$ for $n \in \{1, 2, 3\}$
Require: Distilled reward function $\texttt{reward}_*$
  while stopping criterion not met **do**
    Sample k documents D^k
    for $d \in D^k$ **do**
      for $n \in \{1, 2, 3\}$ **do**
        $\mathcal{C}_d$ = entity clusters with θ_n
        d' = resolve d with $\mathcal{C}_d$
        $\mathcal{T}$ = obtain OpenIE triples for d'
        r = $\text{reward}_n(d')$
        $\hat{g}_k$ = policy gradient for θ_n with reward r
        $\theta_n^{k+1} = \theta_n^k + \alpha_k \hat{g}_k$
      end for
    end for
  end while
  Distilled policy $\theta_* = \frac{\theta_1 + \theta_2 + \theta_3}{3}$
  Sample k documents D^k
  for $d \in D^k$ **do**
    d' = resolve d with $\mathcal{C}_d$
    $\mathcal{T}$ = obtain OpenIE triples for d'
    r = $\text{reward}_*(d')$
    $\hat{g}_k$ = policy gradient for θ_* with reward r
    $\theta_*^{k+1} = \theta_*^k + \alpha_k \hat{g}_k$
  end for
  return Distilled policy θ_*

2.4 EXPERIMENTS

We use a state-of-the-art neural coreference resolution model (Lee et al., 2018a) as our baseline coreference resolver.[3] This model extends Lee et al. (2017a) with coarse-to-fine inference and pre-trained ELMo (Peters et al., 2018).

3 https://github.com/kentonl/e2e-coref

System	Data	Accuracy	F_1 score
RE-KG	12M	0.64	0.78
RE-Text	2M	0.71	0.83
RE-Joint	60K	0.58	0.73
RE-Distill	—	**0.78**	**0.88**

Table 1: Training data size, accuracy and F_1 scores of the reward models on the 200,000 validation triples.

DATA We use the standard training, validation, and test splits from the English OntoNotes.[4] We also evaluate on the English WikiCoref (Ghaddar and Langlais, 2016), with a validation and test split of 10 and 20 documents respectively.

REWARD MODEL TRAINING We use data from English Wikipedia and Wikidata to train our three reward models. For training **RE-KG**, we sample 1 million Wikidata triples, and expand them to 12 million triples by replacing relation phrases with their aliases. For **RE-Text**, we pass the summary paragraphs from 50,000 random Wikipedia pages to Stanford's OpenIE extractor (Manning et al., 2014), creating 2 million triples. For **RE-Joint**, we only use Wikipedia triples that are grounded in Wikidata, resulting in 60,000 triples.[5] We further sample 200,000 triples from Wikidata and Wikipedia for validation, and train the reward models with early stopping based on the F_1 score of their predictions.

EVALUATION All models are evaluated using the standard CoNLL metric, which is the average F_1 over the link-based MUC, entity-based CEAFe, and mention-based B^3 scores (Denis and Baldridge, 2009).

2.5 RESULTS

Since the quality of our reward models is essential to the performance of the coreference resolver adaptations, we first report the validation accuracy and F_1 scores of the four reward models used, in Table 1. We clearly see the advantage of distillation, with a 5% absolute difference between the best single model (**RE-Text**) and **RE-Distill**.

Table 2 presents the downstream effects of applying these reward functions to our baseline coreference policy.[6] The coreference resolution results are similar to the relation extraction results: using a

4 https://catalog.ldc.upenn.edu/LDC2013T19

5 That is, we retain only those triples whose subject and object can be linked to an entity in Wikidata.

6 The models were re-trained from scratch, and the scores are slightly different from those reported in Lee et al. (2018a).

System	OntoNotes	WikiCoref
Lee et al. (2018a)	72.60	57.49
Coref-KG	72.96	57.84
Coref-Text	72.99	57.54
Coref-Joint	72.77	57.51
Coref-Distill	**73.10**	**58.14**

Table 2: Coreference results: average F_1 scores on the OntoNotes and WikiCoref test sets. Differences are significant w.r.t. B^3 (bootstrap test, $p < 0.05$).

distilled policy, learned through multi-task reinforcement learning, leads to better results on both datasets.[7]

While improvements over the current state of the art are relatively small, they reflect significant progress, as they demonstrate the ability to successfully augment coreference resolvers with "free" data from large-scale KB like Wikidata. For relation extraction, this could have positive downstream effects, and also ensure that relations are consistent with real world knowledge. Moreover, this approach has the potential to also be beneficial for coreference resolution in low resource languages, where less annotated data is available, as Wikidata triples are abundant for many languages.

2.6 ANALYSIS

Empirically, we find that fine-tuning the coreference resolver on Wikidata results in two kinds of improvements:

BETTER MENTION DETECTION Since the model is rewarded if the SRO triples produced from the resolved document are present in Wikidata, the model can do well only if it correctly resolves the subject and object, which are usually named entities (more generally, noun phrases). Indeed, we see an improvement in mention detection as exemplified in the first example of Figure 4. Compared to the baseline, the fine-tuned model identifies a larger number of entities, including "southern hemisphere", "Cambridge" and "Oxford", which are missed by the baseline model.

BETTER LINKING As a direct consequence of the above, the model is inclined to also link noun phrases that are not entities. In the second example of Figure 4, we see that "This attempt" is linked to "releasing" by the fine-tuned model. Interestingly, we do not see this type of *eventive* noun phrase linking either in OntoNotes or in the predictions of the baseline model.

7 We repeat this experiment three times with different random seeds and observed the same pattern and very robust performance across the board.

	Baseline system	Fine-tuned system
Mention detection	According to the library's publications, it is the largest academic library in the southern hemisphere. The university has a number of residential college and halls of residence, based on the college system of Cambridge and Oxford universities.	According to the library's publications, it is the largest academic library in the southern hemisphere. The university has a number of residential college and halls of residence, based on the college system of Cambridge and Oxford universities.
Linking	On March 19, Obama continued his outreach to the Muslim world, releasing a New Year's video message to the people and government of Iran. This attempt was rebuffed by the Iranian leadership.	On March 19, Obama continued his outreach to the Muslim world, releasing a New Year's video message to the people and government of Iran. This attempt was rebuffed by the Iranian leadership.

Figure 4: Mention detection and linking examples by the baseline system from Lee et al. (2018a), and the best performing fine-tuned system (Coref-Distill). Mentions of the same color are linked to form a coreference cluster.

This phenomenon, however, also has a side-effect of producing singleton clusters and spurious linking, which adversely affect the recall. On the OntoNotes test data, while the average precision of the best performing fine-tuned model is higher than the baseline (75.62 vs. 73.80), a drop in recall (70.75 vs. 71.34) causes the final F_1 score to only marginally improve.

2.7 RELATED WORK

2.7.1 *Coreference resolution*

Among neural coreference resolvers (Meng and Rumshisky, 2018; Wu and Ma, 2017), Lee et al. (2017a) were the first to propose an end-to-end resolver which did not rely on hand-crafted rules or a syntactic parser. Extending this work, Lee et al. (2018a) introduced a novel attention mechanism for iteratively ranking spans of candidate coreferent mentions, thereby improving the identification of long distance coreference chains. Zhang et al. (2019b) improve pronoun coreference resolution by 2.2 F1 points using linguistic features (gender, animacy and plurality) and a frequency based predicate-argument selection preference as external knowledge. Emami et al. (2018a) incorporate knowledge into coreference resolution by means of information retrieval, finding sentences that are syntactically similar to a given instance, and improving F1 by 0.16.

2.7.2 *Reinforcement learning*

RL has been used for many NLP tasks, including coreference resolution (Clark and Manning, 2016a) and relation extraction (Zeng et al., 2018). Clark and Manning (2016a) use RL to improve coreference resolution by optimizing their mention ranking model and directly use the standard evaluation metrics as the rewards. We, on the other hand, perform end-to-end optimization by rewarding the model's consistency with real world knowledge using relation extraction. To our

knowledge, we are the first to use consistency with world knowledge as a reward for tasks other than knowledge base construction.[8]

2.7.3 *Knowledge bases*

Knowledge bases have been leveraged across multiple tasks across NLP (Bordes et al., 2011; Chang et al., 2014; Lin et al., 2015; Toutanova et al., 2015; Yang and Mitchell, 2017). Specifically for coreference resolution, Prokofyev et al. (2015) implement a resolver that ensures semantic relatedness of resulting coreference clusters by leveraging Semantic Web annotations. Their work incorporates knowledge graph information only in the final stage of the resolver's pipeline, and not during training. In contrast, our work augments information from the knowledge base directly into the training pipeline. Also, they use DBpedia (Auer et al., 2007) as the ontology. Although both Wikidata and DBpedia are designed to support working with Wikipedia articles, DBpedia can be considered as a subset of Wikidata as Wikipedia infoboxes are its main data source. The advantage of Wikidata over DBpedia is its size, and the fact that it is multilingual, which will allow applying our method to other languages in the future.

2.8 CONCLUSION

We presented an architecture for adapting coreference resolvers by rewarding them for being consistent with world knowledge. Using simple multi-task reinforcement learning and a knowledge extraction pipeline, we achieved improvements over the state of the art across two datasets. We believe this is an important first step in exploring the usefulness of knowledge bases in the context of coreference resolution and other discourse-level phenomena. In this area, manually annotated data is particularly expensive, and we believe leveraging knowledge bases will eventually reduce the need for manual annotation.

8 Mao et al. (2018), for example, use reinforcement learning with consistency-like reward to induce lexical taxonomies.

3

ELLIPSIS RESOLUTION AS QUESTION ANSWERING: AN EVALUATION

3.1 ABSTRACT

Most, if not all forms of ellipsis (e.g., '*so does* Mary') are similar to reading comprehension questions ('*what does* Mary do'), in that in order to resolve them, we need to identify an appropriate text span in the preceding discourse. Following this observation, we present an alternative approach for English ellipsis resolution relying on architectures developed for question answering (QA). We present both single-task models, and joint models trained on auxiliary QA and coreference resolution datasets, clearly outperforming the current state of the art for Sluice Ellipsis (from 70.00 to 86.01 F_1) and Verb Phrase Ellipsis (from 72.89 to 78.66 F_1).

3.2 INTRODUCTION

Ellipsis resolution is a hard, open problem in NLP, and an important source of error in machine translation, question answering, and dialogue understanding. There are no large annotated text corpora for this phenomenon, even for English, and we only have annotations for a subset of the known ellipsis constructions. Since annotation is expensive and cumbersome, any synergies with existing NLP tasks could be useful and enable us to leverage auxiliary data when learning models for ellipsis resolution.

This paper presents a simple yet strong approach to ellipsis resolution based on a straightforward observation, depicted in Figure 5, that ellipsis resolution can be converted to a QA problem. Ellipsis and questions put in focus *referentially dependent* expressions (Carlson, 2006), or free variables (Partee, 1978), that need to be resolved in order to comprehend the discourse. For similar observations about different tasks, see McCann et al. (2018a) and Gardner et al. (2019).

This straightforward observation leads us to suggest treating different forms of ellipsis resolution – and later, as an auxiliary task, coreference resolution – as a QA problem, and to apply state-of-the-art architectures for QA to ellipsis resolution tasks, as well as to experiment with using training data for QA and coreference resolution to improve our new ellipsis resolution models.

CONTRIBUTIONS We cast ellipsis as a QA problem, enabling us to induce models for it using neural architectures originally developed for QA. Applying these architectures out of the box enables us to

Sluice Ellipsis

Context: ... But the way things are structured now you have to set aside your ego to make things happen. **The whole thing worked out**. I don't know **how**, but it did. Both sides had to work to make it happen …

Question: I don't know how, but it did.
Answer: The whole thing worked out

Verb Phrase Ellipsis

Context: ... It has to be considered as an additional risk for the investor," said Gary P. Smaby of Smaby Group Inc., Minneapolis. "Cray Computer will be a concept stock," he said. "You either **believe Seymour can do it** again or you **don't** ...

Question: You either believe Seymour can do it again or you don't.
Answer: believe Seymour can do it again

Figure 5: Examples of Sluice Ellipsis and Verb Phrase Ellipsis, represented as "questions" about their associated contexts. Wh-phrases and auxiliary verbs are marked in red and elided phrases are marked in blue.

establish strong results[1] for ellipsis resolution tasks, improving significantly over previous work. Using the same architecture for the different ellipsis resolution tasks, as well as for QA and coreference resolution, enables us to explore synergies between the tasks, and we show that training joint models on these tasks leads to even better performance.

3.3 METHODOLOGY

In this section, we briefly describe the various datasets used for training, and explain how they are converted into QA format. We then move on to the choice of model architectures and the reasoning behind their selection.

3.3.1 *Sluice Ellipsis*

For training and evaluation of Sluice Ellipsis resolution models, we use the corpus introduced by Anand and McCloskey (2015), which contains 3,103 annotated examples of embedded sluices, collected from the New York Times section of the English Gigaword corpus. Since the annotators were free to paraphrase the antecedent, in some cases, a string match on the context does not return antecedent span indices. To ensure a fair comparison, we follow previous work (Rønning et al., 2018), which is also the current state-of-the-art, in ignoring these instances, and use their split for training, development and testing.

1 Though we report state-of-the-art results for both sluice and verb phrase ellipsis, we consider these models as strong baselines for future research as they are obtained purely using existing methods.

Task	Train	Dev	Test	ACL
		ELLIPSIS		
Sluice Ellipsis	1.4k	480	992	351
VP Ellipsis	264	20	78	984
		AUXILIARY		
OntoNotes	153k	18.8k	19.5k	463
WikiCoref	5.6k	630	638	2.2k
SQuAD	87.6k	10.6k	-	117

Table 3: QA pair counts and average context lengths (ACL) for different datasets, after conversion

3.3.2 *Verb Phrase Ellipsis*

Bos and Spenader (2011) provide Verb Phrase (VP) Ellipsis annotations for the WSJ part of the Penn Treebank. All 25 sections were annotated, and we follow them in using sections 0-19 for training, and 20-24 for testing. We further hold out sections 18-19 from the training data for development. This also enables to us compare our results directly with the current state-of-the-art for VP Ellipsis (Zhang et al., 2019d).

3.3.3 *Coreference Resolution*

For coreference resolution, which we use as an auxiliary task, we train and evaluate on two corpora: (i) the English portion of the OntoNotes 5.0[2] corpus with the standard data split used in the CoNLL-2012 shared task (Pradhan et al., 2012a), and (ii) the WikiCoref corpus (Ghaddar and Langlais, 2016), which contains annotations of 30 documents from the English Wikipedia. From this dataset, we use 22 documents for training, 4 documents for development, and 4 for testing.

3.3.4 *QA*

We also use SQuAD v1.1 (Rajpurkar et al., 2016b) as an auxiliary reading comprehension dataset.

3.3.5 *Data Conversion*

For converting the various datasets into the QA format of `<context, question, answer>` triples, we perform a simple restructuring as shown in Figure 5. We consider the entire document as the context; the sentence in which the ellipsis/mention is present becomes the question, and the antecedent/entity becomes the answer. In case of coreference

2 https://catalog.ldc.upenn.edu/LDC2013T19

resolution, where a single sentence can have n mentions, we create n questions where every question is the same sentence with a different mention $i \in \{1 \dots n\}$ marked for resolution with `<ref>` and `</ref>` tags. Table 3 shows the number of QA pairs created from each dataset and the average number of words in their contexts.

3.3.6 *QA Architectures*

Generally, QA models have two main components: (i) an encoder module which learns to represent the question and its context, and (ii) a span selection module which predicts the start and end span indices of the answer if it is present in the context. In this work, we present experiments with three diverse models which take entirely different approaches to build the encoder module: (i) DrQA (Chen et al., 2017), with an LSTM encoder, (ii) QANet (Yu et al., 2018), with a CNN encoder, and (iii) BERT (Devlin et al., 2018), with a (pretrained) transformer encoder. We use the three different models to show that the between-task synergies are relatively robust across architectures; but one architecture (BERT) is clearly superior to the others and will be the standard baseline we propose for future research.[3]

3.4 EXPERIMENTS & RESULTS

We conduct two sets of experiments: (i) the SINGLE-TASK experiments, in which we train and evaluate separate models for the two ellipsis resolution tasks; and (ii) the JOINT modelling experiments, where we train on the best possible combination of ellipsis resolution, coreference resolution and QA data, as determined on the validation set. The results can be seen in Table 4.[4]

SINGLE-TASK SETUP The SINGLE-TASK DrQA model improves the state-of-the-art on sluice ellipsis by 7.48 F_1. The SINGLE-TASK QANet model also improves the state-of-the-art on sluice ellipsis by 5.7 F_1, but fails to learn anything meaningful for VP ellipsis. We hypothesise this is due to the fact that 264 training examples are not enough to train the model's large stack of encoder blocks from scratch.

The SINGLE-TASK BERT model achieves state-of-the-art results in both the ellipsis datasets with absolute error reductions of 50.33% (Sluice Ellipsis) and 13.02% (VP Ellipsis). Interestingly, it also achieves a 17.10% error reduction over the best previously reported results on WikiCoref, but see Appendix A.1.3.2 for why such a direct comparison of numbers is not entirely fair.

3 Note that there are many differences between these architectures; not only the encoder networks. The number of parameters differ, and BERT is pre-trained on large volumes of data. Our purpose here is not comparing strategies, but simply showing that synergies can be seen across all architectures. For more details, see Appendix A.1.2.

4 The reported results are the average of three independent runs with different random seeds.

Task	SotA	Single Task			Joint		
		DrQA	QANet	BERT	DrQA	QANet	BERT
Sluice	70.00	**77.48**	**75.70**	**85.10**	**80.17**	**77.11**	**86.01**
VP	72.89	62.86	1.93	**76.42**	63.54	22.49	**78.66**

Table 4: Ellipsis resolution scores are token-level F_1. Bold-faced results are better than the previous state-of-the-art; underlined results are the new state-of-the-art. When evaluated, our best joint architecture scores 72.31 on OntoNotes and 65.30 on WikiCoref (macro-averages of MUC, B^3, and $CEAF_{\phi_4}$ scores). See Appendix A.1.3.2 for why these numbers are not directly comparable to previously reported coreference resolution results in literature.

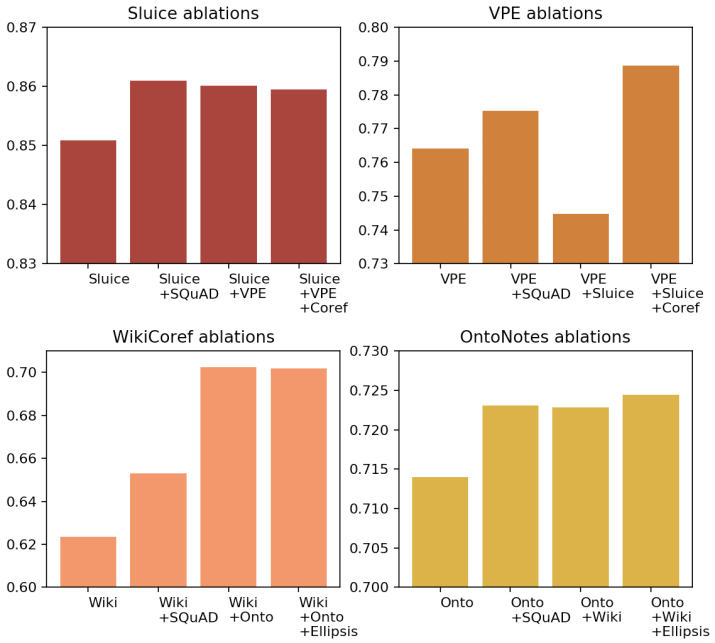


Figure 6: Dataset ablations (F_1)

JOINT SETUP The Joint models always perform on-par with, or better than the Single-Task models. In this setup, the BERT models beat the previous state-of-the-art for both Sluice and VP Ellipsis with 53.37% and 21.28% absolute error reductions respectively.

3.5 DATASET ABLATIONS

We determine the best task combinations on held-out validation data for each ellipsis resolution task.[5] For Sluice Ellipsis, the best results are obtained by training the models on a combination of Sluice and VP Ellipsis data. For VP Ellipsis, the best performance is attained when the models are trained with a combination of all datasets. When training a model for a particular task, we sample auxiliary data from other datasets to match the size of the main task's dataset. For each dataset, the variations in F_1 scores of the best performing architecture

5 These ablations are performed on the best performing (BERT) model.

when combined with other datasets are shown in Figure 6. The most interesting findings from these ablations are mentioned below.

When the two ellipsis datasets are combined, the overall performance of the models increase for both tasks by around 1% each. This shows that the two types of ellipsis are similar, and that when learning ellipsis resolution models, there is considerable synergy between the two resources. If we add subsampled coreference data when training these models, the Verb Phrase Ellipsis models gain up to 2.9%. One possible explanation could be more similarities between noun phrases and verb phrases, than between noun phrases and the sentences that are elided in Sluice Ellipsis resolution.

3.6 ERROR ANALYSIS

We now look at some errors made by our best performing models. First, we compare the errors made by our SINGLE-TASK and JOINT Sluice Ellipsis resolution models before moving on to VP Ellipsis.[6]

3.6.1 *Sluice Ellipsis*

The JOINT Sluice Ellipsis results improve modestly over the SINGLE-TASK Sluice Ellipsis results. This is noteworthy, since the added VP Ellipsis data is quite small compared to the size of the sluice data. These models consistently select an antecedent of the right syntactic form, which is normally a complete sentence. Many of the errors consist of empty outputs: SINGLE-TASK Sluice Ellipsis produces 58 empty outputs, while JOINT Sluice Ellipsis produces 63. Another source of error is discontiguous antecedents. It is not unusual for the gold antecedent to be a discontiguous span (Donecker, 1996), but our models are not permitted to produce such antecedents, so these cases will always be a source of error.

All the systems have problems when the antecedent follows the ellipsis, as in the following example: *I don't know why, but they seem to need a story.* We also compared the right and left periphery scores of sluices, and found better results predicting the right periphery: for SINGLE-TASK Sluice Ellipsis, there were 678 matches on the left edge, and 733 on the right edge; for JOINT Sluice Ellipsis, there were 703 left matches and 734 right matches.

3.6.2 *Verb Phrase Ellipsis*

The SINGLE-TASK VP models trained with just VP Ellipsis data improves on the current state of the art, and further improvement is observed when trained on auxiliary data, especially the Sluice Ellipsis resolution dataset. While the JOINT VP Ellipsis model is generally better than the SINGLE-TASK model, joint training with Sluice Ellipsis resolution data also seems to introduce unfortunate biases. While

6 We also briefly discuss how coreference resolution benefits from synergies with ellipsis in Appendix A.1.3.1.

Context	Then at 10:15, the Dow suddenly started to rebound, and when it shot upward it **did so** even faster than the early-morning fall.	A 190-point drop isn't likely to make much of a dent; multiply that a few times over, though, and **it will**.	Then the whole thing will start to collapse, just as **it did** in the 1970s, and the ghosts and banshees will be howling through the place turning people's hair white.
Gold	shot upward	make much of a dent	collapse
VPE$_s$	shot upward	make much of a dent; multiply that a few times over	go to war to stop anyone from trying to grab Iran. But that ghost wouldn't settle for words, he wanted money and people
VPE$_j$	tt shot upward	a 190-point drop isn't likely to make much of a dent	collapse
	Example (a)	Example (b)	Example (c)

Figure 7: Selected gold and predicted antecedent spans from SINGLE-TASK Verb Phrase Ellipsis (VPE$_s$ in figure) and JOINT Verb Phrase Ellipsis (VPE$_j$ in figure) models.

the SINGLE-TASK model always selects antecedents of the right syntactic form, i.e., verb phrases, the JOINT model may select sentential antecedents. See examples in Figure 7.

In Example (a), the JOINT VP model incorrectly includes the subject *it*, presumably because the sluice data includes complete sentences as antecedents. Similarly in Example (b) – though the SINGLE-TASK model correctly chooses an antecedent beginning with the verb *make*, it continues with additional material that does not form a coherent antecedent. The JOINT result is also incorrect, but note that it consists of the complete sentence containing the correct VP antecedent. Example (b) presents the advantages and disadvantages of the joint ellipsis training data. While the two types of ellipsis require antecedents of different forms, they have similar requirements in terms of where in the context the antecedent is to be found. Example (c) further supports this point. Here the JOINT result is perfect, while the SINGLE-TASK result finds an antecedent that is in the wrong part of the discourse. The SINGLE-TASK model is slightly better with left periphery matches than right: we found 58 left and 55 right matches. This is reversed with the JOINT model, with 54 left and 60 right matches.

3.7 RELATED WORK

We are not the first to use question answering to redefine a set of tasks. Recently, He et al. (2015) showed that semantic role labeling annotations could be solicited by asking simple questions that implicitly target predicate-argument relations in a sentence. Parallel to our work, Hou (2020) cast bridging anaphora resolution as question answering based on context. Wu et al. (2020b) and Li et al. (2020a) also reformulate coreference resolution and named entity recognition as QA. In the realm of re-framing relation extraction as a QA problem,

Levy et al. (2017) and Abdou et al. (2019) create monolingual and multilingual template based QA datasets respectively, which yield relation extraction models which were better at generalizing in the zero-shot setting. Extending this idea, McCann et al. (2018b) introduced the DecaNLP challenge, which casts 10 core tasks in NLP as question-answering problems. Similar to our work, their architecture jointly learns across all of these tasks. DecaNLP includes pronoun resolution, a subset of coreference resolution, but it does so only on a small, hand-crafted dataset; it does not address ellipsis.

LIMITATIONS OF OUR APPROACH One limitation of our approach is that, like most previous work, we assume ellipsis and coreference resolution amount to finding antecedent spans that corefer with the target mention. This is not always the case; the elided material can: (i) have extra-linguistic antecedents, and (ii) refer to something that is contextually implied.

3.8 CONCLUSION

We present strong models for Sluice and Verb Phrase ellipsis resolution problems, by reformulating them as machine reading comprehension problems, significantly outperforming the previously best reported results. We also empirically show that training these models jointly and with auxiliary data from coreference resolution and question-answering further improves their performance. Our code is publicly available at https://github.com/rahular/ellipsis-baselines.

4

JOINT SEMANTIC ANALYSIS WITH DOCUMENT-LEVEL CROSS-TASK COHERENCE REWARDS

4.1 ABSTRACT

Coreference resolution and semantic role labeling are NLP tasks that capture different aspects of semantics, indicating respectively, which expressions refer to the same entity, and what semantic roles expressions serve in the sentence. However, they are often closely interdependent, and both generally necessitate natural language understanding. Do they form a coherent abstract representation of documents? We present a neural network architecture for joint coreference resolution and semantic role labeling for English, and train graph neural networks to model the *coherence* of the combined shallow semantic graph. Using the resulting coherence score as a reward for our joint semantic analyzer, we use reinforcement learning to encourage global coherence over the document and between semantic annotations. This leads to improvements on both tasks in multiple datasets from different domains, and across a range of encoders of different expressivity, calling, we believe, for a more holistic approach to semantics in NLP.

4.2 INTRODUCTION

Coreference resolution and semantic role labeling (SRL) contribute in complimentary ways to forming coherent discourse representations. SRL establishes predicate-argument relations between expressions, and coreference resolution determines what entities these expressions refer to. While often treated separately (He et al., 2018, 2017; Joshi et al., 2019b; Lee et al., 2017b, 2018b), some frameworks consider coreference and semantic roles part of a more holistic meaning representation (Shibata and Kurohashi, 2018). For example, the Groningen Meaning Bank (Bos et al., 2017) annotates documents with discourse representation structures (Kamp and Reyle, 2013), which subsume both levels of analysis; the same holds for other meaning representation frameworks, such as UCCA (Abend and Rappoport, 2013; Prange et al., 2019) and AMR (Banarescu et al., 2013; O'Gorman et al., 2018). However, these frameworks do not offer the simplicity of SRL and coreference annotation, and perhaps consequently require more effort to annotate, and do not have the same amounts of training data (Abend and Rappoport, 2017). Furthermore, comprehensive meaning representation *parsing* approaches (Cai and Lam, 2020; Hershcovich et al., 2017; Liu et al., 2018) tend to be more complex than the sequence tagging or span-based models usually used for coreference resolution and SRL, often referred to as *shallow semantic parsing*.

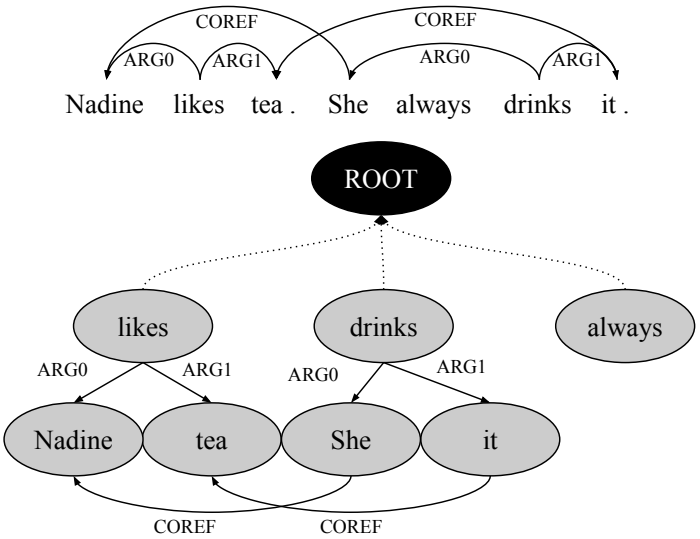


Figure 8: Example coreference and semantic role annotation for a two-sentence document. Top: the original annotation shown as dependencies. Bottom: shallow semantic graph (SSG), where sub-graph heads are connected (with dotted lines) to a dummy root node.

In this paper, we investigate a "minimal" approach to discourse-level semantic parsing, combining coreference and semantic roles in *shallow semantic graphs* (SSGs) that can be seen as a simple, yet rich, discourse-level meaning representations. Consider the two sentences shown in Figure 8, augmented with a (partial) annotation of coreference and semantic roles. A coreference resolver is expected to resolve *Nadine* as an antecedent of *she*, and *tea* as an antecedent of *it*, since these mentions refer to the same entities. A semantic role labeler is expected to detect that these entities are arguments of the predicates *like* and *drink*. The overall semantic analysis corresponds to a coherent and common situation, where someone likes something and consumes it—a very plausible interpretation. This paper presents a model that scores the plausibility or *coherence* of an interpretation based on merged SRL and coreference graphs, or SSGs. While Figure 8 is a simple example that existing SRL and coreference systems will likely handle well, we explore whether such systems in general benefit from feedback from a model that rewards the coherence of their output.

4.2.0.1 *Contributions*

We jointly model coreference resolution and SRL to form discourse-level semantic structures, or SSGs (§4.3). We explicitly model their coherence, presenting a reinforcement learning architecture for semi-supervised fine-tuning of coreference resolvers and semantic role labelers with coherence rewards on unlabeled data (§4.4), improving both coreference resolution and SRL. We present experiments across six encoders of different complexities, six different coreference resolution datasets, and four different SRL datasets (§4.5), showing improvements across all encoders for coreference resolution, and on 4/6 for

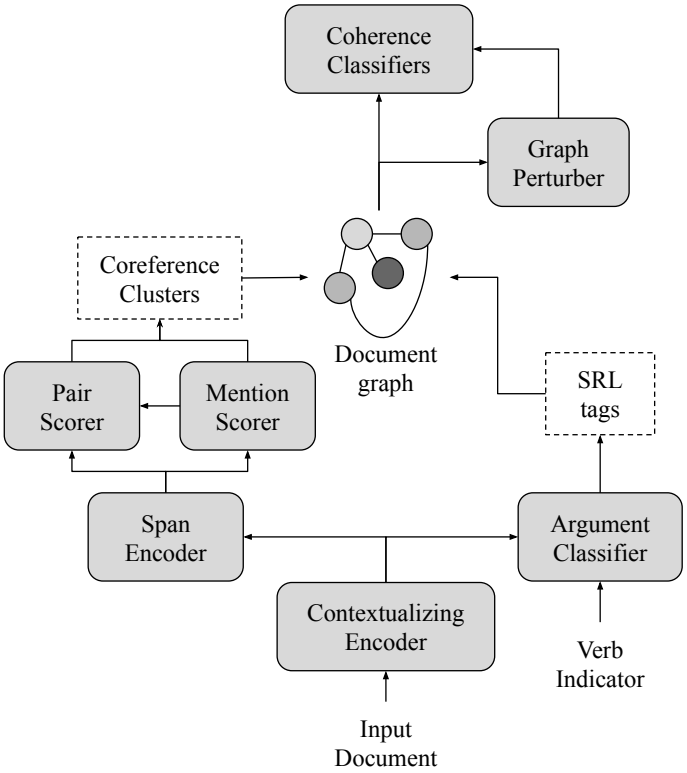


Figure 9: Joint coreference resolution and SRL (bottom half) with a coherence objective (top half). The contextualizing encoder is shared in the multi-task setup, and separate in the single-task one. Predictions from the coreference and SRL models are combined to a document-level SSG, which is scored by coherence classifiers to reward the models.

SRL, for single-task setups; and similar improvements in multi-task setups, where encoder parameters are shared across the two tasks (§6.6). Finally, we analyze the results (§4.8), showing that our fine-tuning setup is particularly beneficial for smaller documents while being on-par with strong baselines on larger documents and that the majority of the remaining coreference errors occur when the antecedent is a pronoun.

4.3 JOINT COREFERENCE RESOLUTION AND SRL

We build baseline single-task and multi-task *supervised models* for coreference resolution and SRL. The overall model architecture is illustrated in Figure 9 (bottom half; till the coreference clusters and SRL tags are generated). In the multi-task setup only the contextualizing encoder is shared. In the single-task setup no parameters are shared.

4.3.1 *Coreference Resolver*

The coreference model is based on the architecture presented in Lee et al. (2017b). Each token's embedding is obtained using a contextualizing encoder. Using a *span encoder*, the token embeddings are

combined into span representations $s(i, j)$, where i and j are the start and end indices in the document. Each span is represented as the concatenation of: (i) its first and last token embeddings, and (ii) an attention-based aggregation of embeddings of all tokens in the span. These span representations are pruned with a *mention scorer*, which outputs the probability of $s(i, j)$ being a coreferent mention. Next, the mention representations are paired together and scored again with a *pair scorer*, which predicts the probability of the mentions referring each other. Coreferring mentions are collected to form clusters. This architecture is combined with pre-trained language models in Lee et al. (2018b) and Joshi et al. (2019b) to get state-of-the-art results.

4.3.2 *Semantic Role Labeler*

The SRL tagger is based on the architecture presented in He et al. (2017). The model uses the contextualizing encoder to embed tokens which are concatenated with a binary indicator to identify whether the token is a verb or not. These token representations are presented to a *argument classifier* for BIO sequence tagging. The current state-of-the-art (He et al., 2018) uses an architecture similar to that of Lee et al. (2017b), where it jointly predicts both arguments and predicates.

4.3.3 *Contextualizing Encoder*

In all setups, we experiment with (i) an LSTM + CNN encoder, and (ii) five BERT (Devlin et al., 2019) encoders of different sizes. In the LSTM + CNN encoder, a bi-LSTM contextualizes words embedded with GloVe (Pennington et al., 2014b) and a CNN encodes individual characters. The final representation is the concatenation of the two. For the BERT encoders, we experiment with different encoder sizes as shown in Table 6, using each token's wordpiece embeddings. Encoder hyperparameters are given in §4.6.

4.4 SEMI-SUPERVISED FINE-TUNING

In the semi-supervised stage of training, classifiers trained on SSGs created from labeled data (Figure 8) are used to fine-tune the supervised models on unlabeled data by reinforcement learning. For each unlabeled document, we use the predicted annotations of the supervised models to build an SSG consisting of SRL predicates and arguments, with links between coreferent mentions. Edge labels are used to distinguish between SRL and coreference edges. These graphs are scored by *graph classifiers* (§4.4.1), trained using graph perturbations (§4.4.2) to model semantic coherence. The confidence value is used as a reward to fine-tune the supervised models using policy gradient (§4.4.3).

Algorithm 2 Training Coherence Classifiers

Require: $\mathbb{G}$: List of SSGs
Require: P: List of perturbations to perform
Require: d: Decay factor
  Initialize clfs = $\emptyset$
  for p in P **do**
    for epoch = 1, ..., N **do**
      Initialize $\mathbb{G}_p = \emptyset$
      for g in $\mathbb{G}$ **do**
        g_p = p (g, d)
        $\mathbb{G}_p$.add (g_p)
      end for
      encoder = DGI ($\mathbb{G}, \mathbb{G}_p$)
      d = decay (d)
    end for
    $data_+$ = (encoder ($\mathbb{G}$), 1)
    $data_-$ = (encoder ($\mathbb{G}_p$), 0)
    clf_p = logistic ($data_+$, $data_-$)
    clfs.add (clf_p)
  end for
  return clfs

4.4.1 *Coherence Classifiers*

We use a graph convolution network (Kipf and Welling, 2017, GCN) to construct continuous representations of the SSGs, where a node representation is composed via a learnt weighted sum of neighboring nodes. Since nodes correspond to text spans, to initialize their representations, we use the supervised model's span encoder. To get the final graph encoding, all the node representations are averaged and compressed using the logistic function as shown in Equation 4.2.

$$\text{graph}_{enc} = \sigma\left(\frac{1}{N}\sum_{i=1}^{N}\text{node}^{i}_{enc}\right) \tag{4.2}$$

The GCN parameters are pre-trained using deep graph infomax (Veličković et al., 2018, DGI), which relies on graph perturbations to learn a task-independent representation. We contrastively train the GCN encoder on gold and *perturbed* graphs, which are generated by randomly perturbing the gold graphs (§4.4.2). We then use the same perturbations to train a logistic regression classifier, with the GCN outputs as features, to discriminate gold graphs from perturbed graphs. As shown in §4.7.2, the trained classifiers are almost perfectly accurate on an unseen development set.

The process for training the coherence classifiers is shown in Algorithm 2. First an SSG $g \in \mathbb{G}$ is built for each labeled document. Then for each type of perturbation $p \in P$, we train one classifier as follows: (i) perturb g to get g_p using perturbation p. We use a decay factor $d \in \{0, 1\}$ to decide the probability of perturbing a sentence in the document. We start with $d = 0.8$ and decay it till $d = 0.1$, (ii)

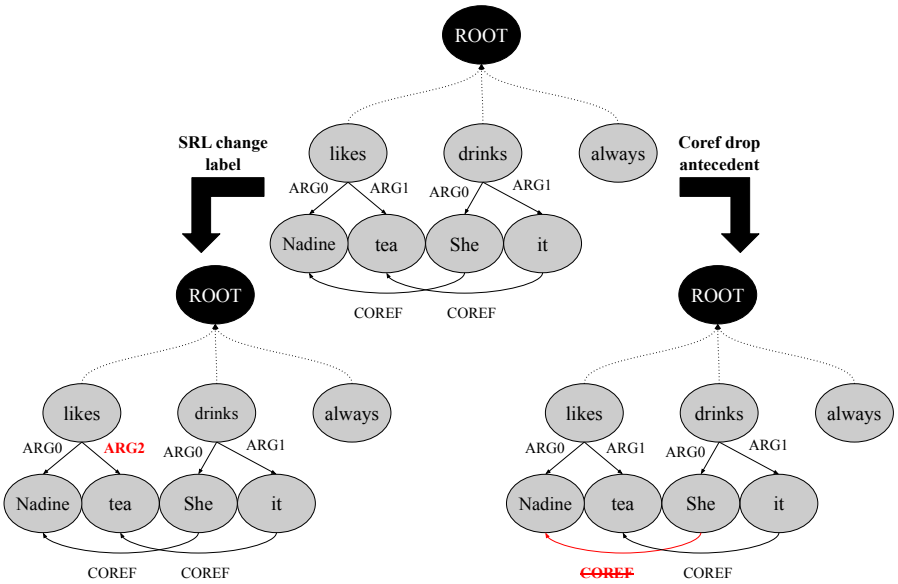


Figure 10: Examples for graph perturbations, starting from the SSG in Figure 8 (center). An 'SRL change label' perturbation is applied to generate a graph (left), where ARG1 is changed to ARG2. A 'Coref drop antecedent' perturbation is applied to generate a graph (right) where a COREF edge is deleted.

once we have a list of perturbed graphs $\mathbb{G}_p$, we train the GCN using DGI, which uses a contrastive loss to learn graph representations such that each pair (g, g_p) is as different to each other as possible, (iii) we use the GCN to get the final representations of graphs in $\mathbb{G}$ and $\mathbb{G}_p$ and create a training dataset consisting of the following (graph, label) pairs: $\{(g, 1) : g \in \mathbb{G}\} \cup \{(g_p, 0) : g_p \in \mathbb{G}_p\}$, and (iv) we train a logistic regression classifier.

4.4.2 *Graph Perturbations*

To train the GCN with DGI, we perturb the gold graphs to reflect the statistics of errors made by the supervised models we want to fine-tune. In general, perturbations are sampled from the following operations: (i) randomly removing edges, (ii) randomly adding edges between existing nodes with a random label, or (iii) randomly adding nodes with a span that is a constituent in the sentence, and a random edge to another existing node. We arbitrarily choose to sample SRL and coreference perturbations with a 3-to-1 ratio.

For SRL perturbations, we rely on the error analysis made by He et al. (2017), whose SRL model is the basis for ours: 29.3% of errors correspond to incorrect argument labels; 4.5% to moved unique arguments; 10.6% to split arguments; 14.7% to merged arguments; 18% to incorrect boundaries; 7.4% to superfluous arguments; and 11% to missed arguments. Consequently, we sample perturbations proportionally to the corresponding error's frequency. We further use He et al. (2017)'s observed confusion matrix of predicted and gold argument labels, sampling replacement labels accordingly. For *coreference*

perturbations, we add a random edge between existing nodes or remove an edge, with uniform probability.

We train one classifier to identify each type of perturbation, resulting in nine different classifiers (seven for SRL and two for coreference; an example for one of each is illustrated in Figure 10). The final confidence for a graph is the average of the individual classifier confidence scores.

4.4.3 *Model Fine-Tuning*

Finally, we use the learned classifiers to fine-tune the underlying coreference resolver and semantic role labeler; using plain text from summary paragraphs of Wikipedia articles, we apply the supervised models to sample an SSG. Using the coherence classifiers' confidence score as a reward, we train the models with policy gradient.

During policy gradient, we consider the selection of SSG edges as actions. More concretely, for coreference resolution, picking the antecedent to each mention is considered an action. Therefore from Figure 8, assuming the model found four mentions ('Nadine', 'tea', 'She', and 'it'), it takes four actions (connecting 'Nadine→ ϕ', 'tea→ ϕ', 'she→Nadine', 'it→tea').[1] For SRL, assigning a label to a token is considered as an action. Therefore the model has to perform nine actions (one for each token) to label Figure 8.

In this work, we assume that all actions are equally good and reward them uniformly. Assigning rewards to individual actions would probably yield better results but is non-trivial and left for future exploration.

4.5 EXPERIMENTS

In this section, we briefly describe the datasets used to train and evaluate our models before moving on to the experimental setup. We then provide implementation details for each stage of the training process and finally present the results of our experiments.

4.5.1 *Datasets*

For supervised training, we use data from the CoNLL-2012 shared task (Pradhan et al., 2012b), which contains data from OntoNotes 5.0[2] with annotations for both coreference resolution and semantic role labeling.

As additional out-of-domain (OOD) development and test data for coreference resolution, we use (i) PreCo (Chen et al., 2018), which contains web-crawled documents and data from the RACE dataset (Lai et al., 2017); (ii) Phrase Detectives (Poesio et al., 2013a), which contains two evaluation sets, one sampled from Wikipedia and the other from the Gutenberg project; (iii) WikiCoref (Ghaddar and Langlais, 2016),

1 ϕ indicates no antecedent

2 https://catalog.ldc.upenn.edu/LDC2013T19

Hyperparameters	Lee et al. (2018)	Joshi et al. (2019b)	Ours
max. span width	30	30	10
cxt. enc. (layers/dims)	3/1024	24/1024	12/768*
span enc. (layers/dims)	3/400	-	1/400
pruner (layers/dims)	2/150	1/1000	1/150
top span ratio	0.4	0.4	0.3
max antecedents	250	50	100
course to fine inference	True	True	False

Table 5: Comparison of hyperparameters between state-of-the-art and our coreference models. *This value is for BERT-Base. See Table 6 for other sizes.

which contains long form documents from the English Wikipedia; and (iv) WinoBias (Zhao et al., 2018), which is focused on gender bias with Winograd-schema style sentences, authored manually.

For SRL, we additionally use (i) the CoNLL-2005 shared task data (Carreras and Màrquez, 2005), which contains two evaluation sets: the in-domain WSJ set and the OOD Brown set; and (ii) English Web Treebank (Silveira et al., 2014)[3], which contains weblogs, newsgroups, email, question-answers and review text.

4.5.2 *Experimental Setup*

We first train the coreference and SRL models (§4.3) using supervised learning, and the coherence classifiers on gold graphs and their perturbations. Both are trained on the CoNLL-2012 training set. We then fine-tune the models by semi-supervised learning (§4.4), with the summary paragraphs of 10,000 randomly sampled English Wikipedia articles.[4] We test our models across six domains for coreference resolution, and four domains for SRL, using in-domain evaluation data.

4.6 IMPLEMENTATION DETAILS

Since the goal of this work is not to surpass the state of the art, but to demonstrate that discourse-level coherence can be used to improve shallow semantic analysis, and due to memory and compute constraints, we use smaller versions of the best performing architectures in the literature as baselines.

4.6.1 *Coreference Model*

We use the same architecture that state-of-the-art coreference systems like Lee et al. (2017b, 2018b) and Joshi et al. (2019b) use, but with lesser capacity. A comparison of the important hyperparameters that

3 https://catalog.ldc.upenn.edu/LDC2017T15

4 https://www.wikipedia.org, dump from March 4, 2019.

Encoder	# layers	dim
LSTM + CNN	1	500
BERT-Tiny	2	128
BERT-Mini	4	256
BERT-Small	4	512
BERT-Medium	8	512
BERT-Base	12	768

Table 6: Number of layers and the output dimension of our contextualizing encoders.

vary between our model and the current state-of-the-art is shown in Table 5.

4.6.2 *SRL Model*

He et al. (2017) use 8 LSTM layers with highway connections and recurrent dropout. We replace this encoder with each of our contextualizing encoder configurations. Following He et al. (2017), we also use constrained decoding to produce only valid BIO tags as output.

4.6.3 *Contextualizing Encoders*

For the LSTM + CNN encoder, 300-dimensional GloVe embeddings (Pennington et al., 2014b) are fed into a bi-LSTM with a hidden size of 200, to get a 400-dimensional word representation. We concatenate this with 100-dimensional character embeddings obtained from a CNN character encoder with a filter size of 5. The other five encoders are based on the standard BERT recipe (Turc et al., 2019), and their sizes can be seen in Table 6.

4.6.4 *Supervised Training*

For training both single-task and multi-task models, we use the Adam optimizer (Kingma and Ba, 2014) with a weight decay of 0.01 and initial learning rate of 10^{-3}. For BERT parameters, the learning rate is lowered to 10^{-5}. We reduce the learning rates by a factor of 2 if the evaluation on the development sets does not improve after every other epoch. The training is stopped either after 100 epochs, or when the minimum learning rate of 10^{-7} is reached. In the multi-task setup, we sample a batch from each task with a frequency proportional to the dataset size of that task. All experiments are run on a single GPU with 16GB memory. The hyperparameters were manually selected to accommodate for training time and resource limitation, and were not tuned based on model evaluation.

	Perturbation type	Accuracy (%)
SRL	change label	98.98
	move argument	99.88
	split spans	99.72
	merge spans	99.29
	change boundary	98.96
	add argument	99.22
	drop argument	100.00
Coref	add antecedent	99.10
	drop antecedent	100.00

Table 7: Graph classifier development accuracy.

4.6.5 *Coherence Classifiers*

The GCN encoder used to encode the SSGs has 512 hidden channels and is trained with Adam for 10 epochs. We use a 20-dimensional embedding to represent the type of node and a binary indicator to represent the edge type.

4.6.6 *Finetuning*

The supervised models are fine-tuned for 10 epochs with the same optimizer configuration. Only the learning rate is changed to $3 \cdot 10^{-4}$. Hill climbing is used during policy gradient, i.e., if fine-tuning on a batch of Wikipedia documents does not yield an improvement, the parameters are reset to their previous best state.

In the multi-task setup, the coreference resolution and SRL sub-models are fine-tuned separately. This is because we do not want to sample actions for both tasks as it makes the constructed SSG more noisy. For constructing the SSGs in the single-task setup, we use the best performing SRL model for fine-tuning the coreference resolution model, and the best performing coreference resolution model for fine-tuning the SRL model.

4.7 RESULTS

4.7.1 *Coreference Resolution and SRL*

The mean F_1 over MUC, $CEAF_{\phi_4}$, and B^3 scores averaged across the six test sets for coreference resolution and the macro-averaged F_1 scores of the four test sets for SRL (including in-domain and out-of-domain), for each of the six encoder configurations, is presented in Table 8. The individual results for each dataset are presented in the Appendix; Tables 18 and 20 for single-task models, and in Tables 19 and 21 for multi-task models respectively.

We see substantial improvements from coherence fine-tuning across the board for all coreference tasks. Results for single-task SRL improves in all settings except for BERT-mini and BERT-medium encoders. In the multi-task setting for SRL, we see consistent improvements with two exceptions: the results for LSTM + CNN and BERT-base. Coreference resolution generally improves more than for SRL.

4.7.2 *Coherence Classifiers*

The accuracy of the nine coherence classifiers (§4.4.1) on the CoNLL-2012 development set is shown in Table 7, which indicate that the classifiers can almost perfectly detect perturbed graphs, and shows their effectiveness at providing a reward signal to the models. While it could be argued that the perturbations are too easy to detect, observing the perturbed graphs (exemplified in Figure 10) leads to the impression that they require sensitivity to distinctions that are important for correct coreference resolution, SRL and the coherence between them. Indeed, the rewards lead to improvements in each of the tasks.

Encoder	Single-Task				Multi-Task			
	Coreference		SRL		Coreference		SRL	
	Base.	Ours	Base.	Ours	Base.	Ours	Base.	Ours
LSTM + CNN	49.01	**49.40**	67.63	**67.74**	48.65	**49.60**	**67.28**	67.05
BERT-**Tiny**	49.70	**50.95**	56.87	**57.08**	45.65	**51.17**	56.65	**56.85**
BERT-**Mini**	52.61	**52.88**	**70.51**	70.48	50.14	**53.02**	71.10	**71.13**
BERT-**Small**	52.76	**53.90**	74.26	**74.48**	51.26	**53.73**	74.72	**74.77**
BERT-**Medium**	55.67	**56.19**	**75.62**	75.57	51.48	**55.52**	77.89	**78.01**
BERT-**Base**	57.78	**58.18**	79.46	**79.52**	56.40	**57.55**	**80.25**	80.19

Table 8: COREFERENCE RESOLUTION and SEMANTIC ROLE LABELING results of single-task and multi-task models. 'Base.' and 'Ours' represent the the supervised baseline and coherence fine-tuned models respectively. The numbers are the mean of MUC, $B^3$and $CEAF_{\phi_4}$(macro-averaged) F_1 scores averaged over six (four) coreference (SRL) datasets.

4.8 ERROR ANALYSIS

By analysing the results of the fine-tuned models on all datasets (Table 8), we make the following observations:[5]

4.8.1 *Document length*

Fine-tuning leads to larger improvements on smaller documents (see Figure 11). This is likely because the unlabeled data we use for fine-

5 Unless mentioned otherwise, all analysis is carried out on the single-task BERT-Base model.

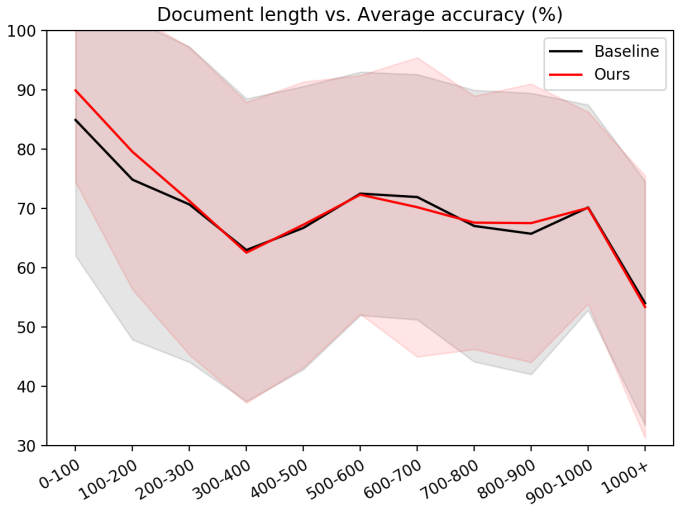


Figure 11: Percentage of correct predictions of our BERT-Base coreference model across all datasets plotted against document lengths.

tuning consists of short paragraphs. While using longer documents for fine-tuning was not possible due to memory constraints, we expect that this will increase the model's sensitivity to long-distance inter-dependencies, and further improve its performance on these documents.

4.8.2 *Coreference resolution vs. SRL*

In general, SRL sees smaller improvements from fine-tuning with policy gradient than coreference resolvers, probably because it is harder to assign credit to specific model decisions (Langford and Zadrozny, 2005). Semantic role labeling of a paragraph typically requires a much longer sequence of actions than determining coreference, leading to limited benefit from reinforcement learning. Similar results have been observed in machine translation (Choshen et al., 2020).

4.8.3 *Precision vs. recall*

Precision often increases after fine-tuning whereas recall decreases. Similar effects have been reported for knowledge-base grounding of coreference resolvers (Aralikatte et al., 2019b).

4.8.4 *Encoder sizes*

From the results, we also see that our fine-tuning approach is robust to encoder sizes with improvements across the board. It is particularly interesting to see that the multi-task BERT-Tiny coreference models come close or even surpass the bigger BERT-Base models on datasets like PreCo and WinoBias, which contain short documents (see Table 19 in the Appendix).

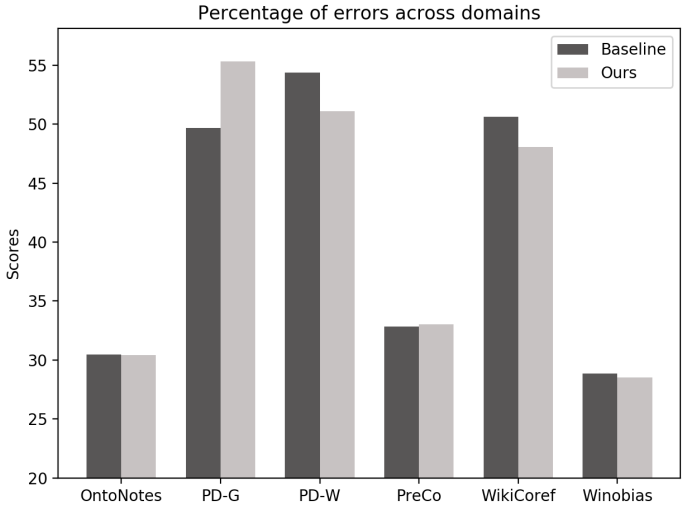


Figure 12: Percentage of errors over the total number of predictions that our coreference system makes across each domain of the evaluation data.

In both single-task and multi-task setups, fine-tuning helps the smaller coreference models more than the larger ones, which are already more accurate. This trend is expected as the larger models tend to be over-parameterized.

4.8.5 *Domain adaptation*

We also perform an error analysis to identify the domains which are hard for our coreference models (see Figure 12). We find that our coherence fine-tuned model always performs better than or on par with the supervised baseline model, expect in the case of Phrase Detectives - Gutenburg (PD-G). We postulate that the increase in PD-G errors can be attributed to the length of the documents in the dataset.[6]

4.8.6 *Part-of-speech*

As seen in Figure 13, across all domains, most errors from the coherence fine-tuned system occur when the antecedent is a pronoun, except for WikiCoref, where the most errors occur when the antecedent was a multi-word expression. This trend is seen in the supervised baseline models as well.

Apart from being the most frequent among mentions, two possible reasons why pronouns could be predicted incorrectly most often are: (i) as the distance in text increases between the original antecedent and subsequent pronouns, it becomes more difficult to resolve, and (ii) as a text becomes more complex, with multiple possible antecedents to choose from, linking becomes harder. Given the increased performance of our coreference resolver from the inclusion of

6 The average document length of PD-G is 1507.2 tokens, which is the highest among all datasets.

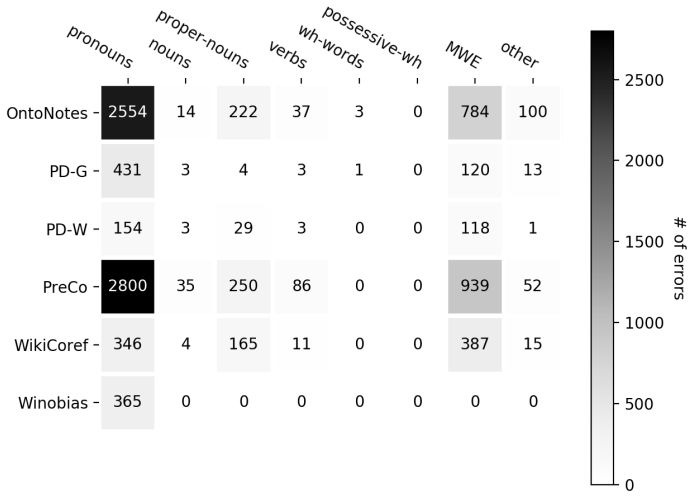


Figure 13: Heatmap showing the POS-tag categories for the antecedents that our fine-tuned coreference system incorrectly classified. All domains except WikiCoref have the highest amount of errors made when the antecedent is a pronoun. Here, pronouns are PRP, PRP$; MWE is any multi-word expression, nouns are NN, NNS; proper-nouns are NNP, NNPS; verbs are VB, VBD, VBG, VBN, VBP, VBZ; other tags we observed were IN, JJR, JJ, RB, DT, CD, MD, POS; and wh-words are WDT, WRB, WP, WP$.

a coherence classifier, we hypothesize that the second problem would be easier for our system to overcome, while the first could still persist.

4.8.7 *Span length*

Finally, we analyse the length of the mentions linked by our models. In general, both supervised baseline and coherence fine-tuned models perform similarly for very short (0–3 tokens) and very long (7+ tokens) mentions. However, we see an improvement in linking accuracy of the coherence fine-tuned model when the mention length is between 3–7.

4.9 RELATED WORK

4.9.1 *Augmented Coreference Resolution*

Previous work has augmented Coreference resolvers with syntax information (Clark and Manning, 2016b,d; Wiseman et al., 2016), external world knowledge (Aralikatte et al., 2019b; Emami et al., 2018b; Rahman and Ng, 2011b) and a variety of other linguistic features (Haghighi and Klein, 2009; Ng, 2007; Zhang et al., 2019a). Similarly, Ponzetto and Strube (2006a,b) used features from SRL and external sources for a non-neural coreference resolver.

4.9.2 *Augmented Semantic Role Labelling*

SRL systems have long utilised annotations from syntactic formalisms as an essential component (Hacioglu, 2004; Levin, 1993; Pradhan et al., 2005; Punyakanok et al., 2008; Sutton and McCallum, 2005). More recently, Strubell et al. (2018) showed that it was possible to exploit information from syntactic parses for supervision of the self-attention mechanism in a fully differentiable transformer-based SRL model, surpassing the previous state-of-the-art. Xia et al. (2019) follow up on this, presenting a detailed investigation into various methods of incorporating syntactic knowledge into neural SRL models, finding it consistently beneficial.

4.9.3 *Document Level Consistency*

Document-level modelling has been shown to be beneficial for NLP tasks such as machine summarization (Chen et al., 2016), translation (Junczys-Dowmunt, 2019; Maruf and Haffari, 2018; Voita et al., 2018), sentiment analysis (Bhatia et al., 2015), and question answering (Sadek and Meziane, 2016; Verberne et al., 2007). For semantic analyzers, document-level consistency is an important requirement. Indeed, when training on complete documents, it also provides a strong input signal. In previous work Tang et al. (2015) presented a user product neural network and validated the effects of users and products in terms of sentiment and text-based consistency. Likewise, Du et al. (2019) used label consistency as an additional objective for a procedural text comprehension model, showing state-of-the-art performance. More recently, Liu and Lapata (2018) used discourse structure and global consistency to guide a machine comprehension model.

Our approach is orthogonal and possibly complementary to those described above: we investigate the consistency in the overall information presented in complete documents for span graphs composed of semantic role labeling and coreference resolution annotations.

4.10 CONCLUSION

We presented a joint coreference resolver and semantic role labeler along with a method of fine-tuning them with document-level coherence rewards over unlabeled documents. We find that this leads to considerable performance gains for coreference resolution across domains, and moderate improvements for semantic role labeling. Results are presented across six English coreference resolution datasets and four English semantic role labeling datasets. Our code will be made publicly available at `https://github.com/rahular/joint-coref-srl`

5

MODEL-BASED ANNOTATION OF COREFERENCE

5.1 ABSTRACT

Humans do not make inferences over texts, but over *models* of what texts are about. When annotators are asked to annotate coreferent spans of text, it is therefore a somewhat unnatural task. This paper presents an alternative in which we preprocess documents, linking entities to a knowledge base, and turn the coreference annotation task – in our case limited to pronouns – into an annotation task where annotators are asked to assign pronouns to entities. Model-based annotation is shown to lead to faster annotation and higher inter-annotator agreement, and we argue that it also opens up for an alternative approach to coreference resolution. We present two new coreference benchmark datasets, for English Wikipedia and English teacher-student dialogues, and evaluate state-of-the-art coreference resolvers on them.

5.2 INTRODUCTION

Language comprehension is often seen as the incremental update of a mental model of the situation described in the text (Bower and Morrow, 1990). The model is incrementally updated to represent the contents of the linguistic input processed so far, word-by-word or sentence-by-sentence. In this paper, we restrict ourselves to one central feature shared by most theories of mental models: they include a list of entities previously introduced in the text. This corresponds to the *constants* of first-order models or the referents associated with different *roles* in frame semantics. By models we thus simply mean a set of entities. Obviously, this is not sufficient to represent the meaning of texts, but focusing exclusively on annotating nominal coreference, we can ignore relations and predicates for this work. We will use the term *model-based annotation* to refer to linguistic annotation using model representations to bias or ease the work of the annotators.

Mental models have previously been discussed in linguistics literature on coreference (Runner et al., 2003). The motivation has often been that some pronouns refer to entities that are not explicitly mentioned in the previous text, but are supposedly available in the reader's mental model of the text, by inference. Consider, for example:

(1) I knocked on the door of room 624. *He* wasn't in.

The introduction of the referent of *he* in (1) is implied by the introduction of the entity *room 624*. In this paper, we present a new approach to annotating coreference that enables simple annotation of examples such as (1): Instead of asking an annotator to relate pronouns

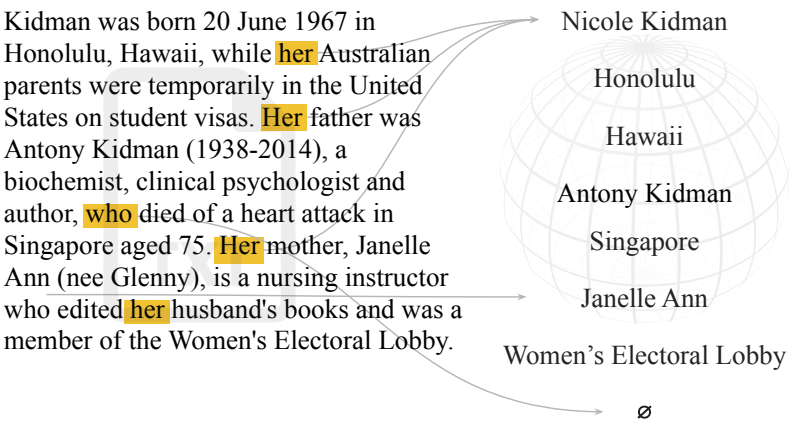


Figure 14: Example of an annotation from the dataset.

and previous spans of text, we ask the annotator to link pronouns and entities in document models. Moreover, we argue that model-based annotation reduces the cognitive load of annotators, which we experimentally test by comparing inter-annotator agreement and annotator efficiency across comparable annotation experiments. Fig. 14 showcases a concrete example from the collected dataset.

CONTRIBUTIONS This paper makes a technical contribution, a conceptual contribution, and introduces a novel corpus annotated with coreference to the NLP community: (a) The technical contribution is a novel annotation methodology, where annotation is mediated through a model representation. We believe similar techniques can be developed for other NLP tasks; see §6 for discussion. (b) The conceptual contribution is a discussion of the importance of mental models in human language processing, and an argument for explicitly representing this level of representation in NLP models. (c) Our corpus consists of manually annotated sentences from English Wikipedia and QuAC (Choi et al., 2018). In addition to the model-based annotations, we also provide the coreference links obtained in our baseline experiments.

5.3 RELATED WORK

5.3.1 *Annotation interfaces*

The idea of easing the cognitive load of annotators by changing the way data is represented, is at the core of many papers on annotation interfaces. Early tools like MMAX2 (Müller and Strube, 2006) provide a clean user interface for annotators by highlighting mentions and connecting entity chains to visualize coreference along with helpful features like inter-annotator agreement checker, corpus querying, etc. Newer tools like WebAnno (Day et al., 2004; Yimam et al., 2013) ease the process of annotation by having support for flexible multi-layer annotations on a single document and also provide project management utilities. APLenty (Nghiem and Ananiadou, 2018) provides au-

tomatic annotations for easing annotator load and also has an active learning component which makes the automatic annotations more accurate over time.

For relieving annotator load, these tools form clusters of coreference such that the annotator can choose to link a mention to one of these clusters. But this is possible only after the clusters are well-formed i.e. after some amount of annotation has taken place. One advantage of our approach is that we provide representatives for each cluster (the entities in the document) right from the start of the annotation process.

5.3.2 *Mental models in NLP*

Culotta et al. (2007) present a probabilistic first-order logic approach to coreference resolution that implicitly relies on mental models. Peng et al. (2015) focus on hard Winograd-style coreference problems and formulate coreference resolution as an Integer Linear Programming (ILP) to reason about likely models. Finkel and Manning (2008) also explore simple ILPs over simple first-order models for improving coreference resolution. They obtain improvements by focusing on enforcing transitivity of coreference links. In general, the use of first order models has a long history in NLP, rooted in formal semantics, going back to Fregean semantics. Blackburn and Bos (2005), for example, present a comprehensive framework for solving NLP problems by building up first order models of discourses.

5.3.3 *Coreference datasets*

The main resource for English coreference resolution, also used in the CoNLL 2012 Shared Task, is OntoNotes (Pradhan et al., 2012a). OntoNotes consists of data from multiple domains, ranging from newswire to broadcast conversations, and also contains annotations for Arabic and Chinese. WikiCoref (Ghaddar and Langlais, 2016) is a smaller resource with annotated sentences sampled from English Wikipedia. Our dataset includes paragraphs from all pages annotated in WikiCoref, for comparability with this annotation project. See §5 for discussion. Several other coreference datasets have been introduced recently: GAP (Webster et al., 2018) is another evaluation benchmark, also sampled from Wikipedia and focuses on addressing gender bias in coreference systems. Phrase Detectives (Poesio et al., 2013b) gamifies the creation of anaphoric resources for Wikipedia pages, fiction and art history texts. Cohen et al. (2017) annotate journal articles to create the CRAFT dataset which has structural, coreference and concept annotations. The annotation process of this dataset is similar in spirit to ours as their concept annotations link text mentions to curated ontologies of concepts and entities.

5.4 DATA COLLECTION

We collect 200 documents[1] from two sources: (i) the summary paragraphs of 100 English Wikipedia documents (30 titles from WikiCoref and 70 chosen randomly), and (ii) the first 100 datapoints from the Question-Answering in Context (QuAC) dataset. Every QuAC document contains a Wikipedia paragraph and QA pairs created by two annotators posing as a student asking questions and a teacher answering the questions by providing short excerpts from the text. Thus the domain of all the documents is English Wikipedia.

5.4.1 *Design Decisions*

Some Wikipedia articles have short summaries with very few pronouns and some do not have summaries at all. Therefore, for each document chosen randomly, we first verify if it has a summary that contains at least five pronouns. If it does not, we choose another document and repeat this process till we get the required number of documents. We then extract all the entities from every document by parsing URL links present in the document which link to other Wikipedia pages or Wikidata entities. For QuAC documents, where all links are scrubbed, we parse their original Wikipedia pages to get the entities. Lastly we remove all markups, references and lists from the documents.

We collect a comprehensive list of English pronouns for linking. Some pronouns by their definition, almost never refer to entities. For example, (i) interrogative pronouns: 'what', 'which', etc., (ii) relative pronouns: 'as', 'who', etc., and (iii) indefinite pronouns: 'anyone', 'many', etc. For completeness, we do not remove these words from the list. We however allow the annotators to mark them specifically as *No Reference*.

5.4.2 *Annotation*

To test our hypothesis that model-based coreference annotations are faster to create and more coherent, we pose two tasks on Amazon Mechanical Turk (AMT): (i) *Grounded task*: where all the parsed entities from a document are displayed to the annotator for linking with the pronouns, (ii) *Span annotation task*: where the entities are not shown and the annotator is free to choose any span as the antecedent. 30 documents from each source are doubly annotated to compute the inter-annotator agreement and the other 70 were singly annotated.

An annotation tool with two interfaces is built, one for each task, with slight differences between them as shown in Figures 15 and 16 respectively. The tool takes in a pre-defined list of mentions (pronouns in our case) which are markable. The annotators can link only these words with coreferent entities. This reduces the cognitive load on the

1 We use the term *document* to denote a datapoint in our dataset.

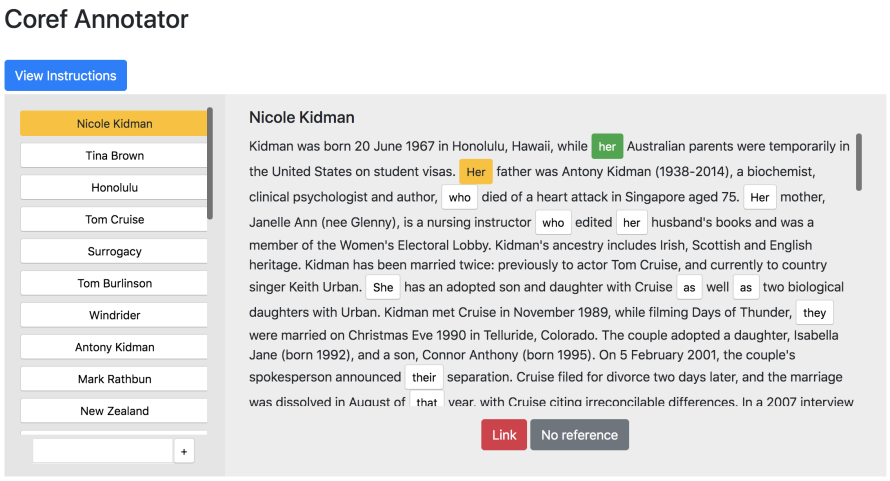


Figure 15: Screen grab of the interface for the grounded-annotation task

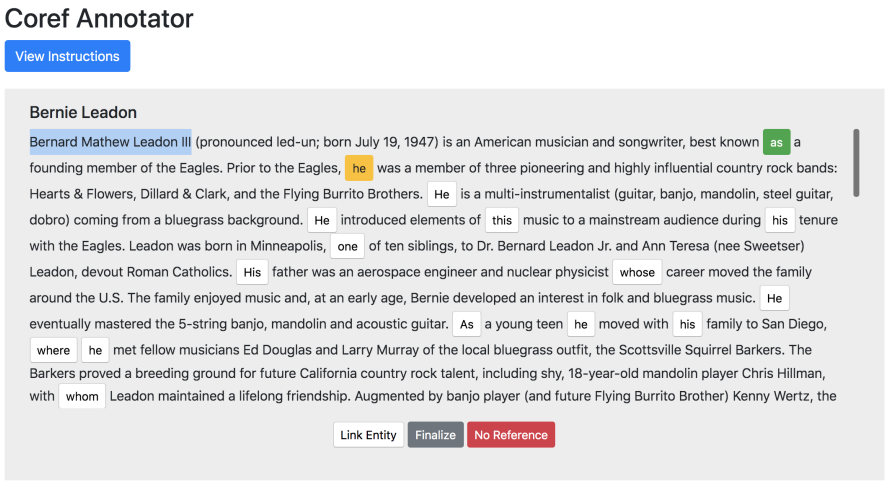


Figure 16: Screen grab of the interface for the span-annotation task

annotators. The annotation process for the two tasks is briefly described below.

GROUNDED TASK For this task, the interface (Fig. 15) is split into two parts. A larger part on the right contains the document text and the mention pronouns are highlighted in white. A sidebar on the left is populated with all the entities extracted from the document. In case of missing entities, the annotator has the option of adding one using the input box present at the bottom-left of the screen. The annotators are asked to link the mention pronouns in the document with one or more entities by: (i) clicking on a mention, (ii) clicking on one or more entities; , and (iii) clicking on the red *Link* button. If any mention does not have an antecedent, the annotators are asked to mark them with the grey *No reference* button. The color of the currently selected mention and entities are changed to yellow for convenience. Mentions which are already annotated are marked in green.

SPAN ANNOTATION TASK In this task, the interface does not have the sidebar (Fig. 16) and the annotators are free to mark one or more

spans in the document as the antecedent(s) for a mention pronoun by selecting the span(s) with their pointers. In a scenario where one mention pronoun has to be linked with multiple antecedents, the annotators have to highlight the spans and click on the white *Link Entity* button multiple times. Therefore, an additional red *Finalize* button is provided to mark the end of one linking episode. Apart from the lack of the entity sidebar and inclusion of the previously mentioned *Finalize* button, all other features of the interface remain the same as those for the Grounded task.

AMT DETAILS The annotation tasks were open only to native English speakers whose approval rate was above 90% and they had ten minutes to annotate a document. Every fifth document annotated by an annotator was a secret test document for which annotations were known. The annotators were allowed to continue only if there was more than 90% match between the gold and their annotations. Each task was published 15 days apart to diversify the annotator pool.

5.5 EXPERIMENTS

5.5.1 *Inter-annotator agreement*

As mentioned in Section 5.4.2, we doubly annotate 30 documents from each source to measure the inter-annotator agreement and the results are presented in Table 9. The numbers clearly indicate that the grounded tasks introduce less uncertainty about the antecedents and hence result in more agreements between the annotators. Ideally the exact match and F_1 scores for grounded tasks should be identical. However, the slight difference observed is because of mentions being linked to different, but similar looking entities. For example, in the sentence "Harry Potter is a global phenomenon. *It* has captured the imagination of . . . ", the mention *It* can be linked either to Harry Potter – the movies or Harry Potter – the books.

5.5.2 *Annotation times*

We can estimate the cognitive load on the annotators by measuring the time taken for marking the documents. Figure 17 shows the mean annotation times and their standard deviations for annotating documents in different settings. In general, QuAC documents require more time and effort to annotate due to the presence of QA pairs which require the annotators to possibly re-read a portion of the context paragraph. Also, it is clear that grounding the document eases the load on annotators irrespective of the source of documents.

5.5.3 *State-of-the-art*

We run our data through three state-of-the-art coreference resolution systems and report the average precision, recall and F_1 scores of three

	Exact Match	F_1 Score
Wiki grounded	0.70	0.74
Wiki free	0.50	0.65
QuAC grounded	0.65	0.67
QuAC free	0.52	0.64

Table 9: Inter-annotator agreement scores

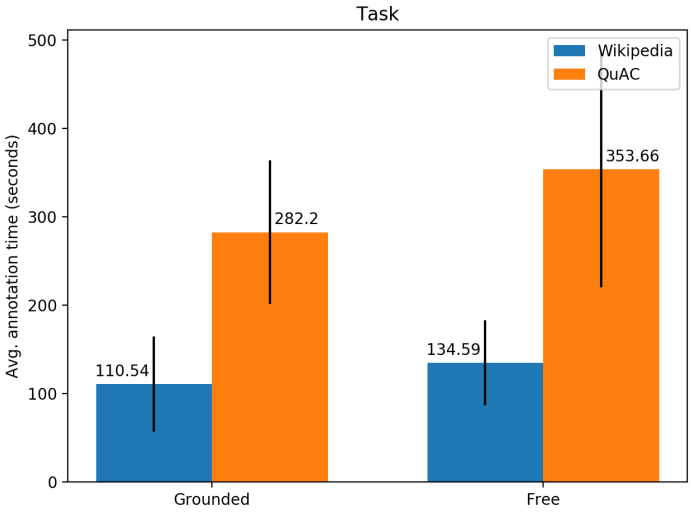


Figure 17: Average annotation times for the two tasks and settings

standard metrics: MUC, B^3 and CEAFe (Cai and Strube, 2010), in Table 10.[2] While Clark and Manning (2016c)[3] and Lee et al. (2018c) train on OntoNotes 5 to perform both mention detection and linking, Aralikatte et al. (2019a) use a multi-task architecture for resolving coreference and ellipsis posed as reading comprehension, which is also trained on OntoNotes 5, but uses gold bracketing of the mentions and performs only mention linking.[4] The results show that the dataset is hard even for the current state-of-the-art and thus a good resource to evaluate new research.

5.6 DISCUSSION

The main purpose of this work is to study how humans annotate coreference with and without grounding. Therefore we give freedom to the annotators by asking them to abide by a minimal set of rules. We see interesting annotation patterns in our dataset: Generally, the indefinite pronoun 'all' is marked as having 'No Reference'. But for

2 Converting our grounded data to the OntoNotes format is in some cases lossy, since entity aliases may not perfectly match previous mentions.

3 We use an improved implementation available at https://github.com/huggingface/neuralcoref.

4 This explains the comparatively higher numbers. See discussion in their paper for more details.

System	Wiki			QuAC		
	P	**R**	**F_1**	**P**	**R**	**F_1**
Clark and Manning (2016c)	24.72	32.87	27.95	20.15	27.98	23.39
Lee et al. (2018c)	21.38	37.90	26.67	17.42	39.07	23.79
Aralikatte et al. (2019a)*	43.88	48.58	45.96	46.18	46.23	46.14

Table 10: The macro-averages of MUC, B^3, and $CEAF_{\phi_4}$. (*assumes gold brackets for mentions.)

the sentence "... Harry Potter, and his friends Hermione Granger and Ron Weasley, *all* of whom ...", for example, the pronoun 'all' is linked as follows: (i) in the grounded task, the word is linked to three entities – Harry Potter, Hermione Granger and Ron Weasley, whereas (ii) in the span annotation task, the word is linked to the phrase "Harry Potter, and his friends Hermione Granger and Ron Weasley". We see that the annotation for the grounded task is cleaner than that for the span annotation task. This effect is observed throughout the dataset. Also, in span annotation tasks, while some annotators link mention pronouns to the first occurrence of an entity, some link them to the latest occurrence, sometimes resulting in multiple clusters instead of one. By design, this is not the case in the grounded tasks.

5.6.1 *Comparison with WikiCoref*

WikiCoref has 30 annotated pages from English Wikipedia. Our dataset contains 200 documents of which 30 titles are the same as those of WikiCoref. WikiCoref uses the full Wikipedia page for annotation, whereas we extract only the summary paragraphs from each page. WikiCoref doubly annotates only 3 documents for reporting inter-annotator agreement, whereas we do it for 30 documents. The inter-annotator agreements themselves are not comparable because they only report the Kappa coefficient for mention identification which does not occur in our tasks.

5.6.2 *Generalization to other NLP tasks*

Our first annotation experiments have been limited to coreference for pronouns, but obviously the same technique can be used to annotate other linguistic phenomena involving relations between noun phrases, e.g., other forms of coreference, nominal ellipsis, implicit arguments, or roles of semantic frames. Our models only include individuals or constants, but if we extend our models to also include propositions holding for individuals or between individuals, we could potentially also do grounded annotation of complex verbal phenomena such as VP ellipsis, gapping, sluicing, etc.

5.7 CONCLUSION

We propose a new way of annotating coreference by grounding the input text to reduce the cognitive load of the annotator. We do this by making the annotators choose the antecedent for mentions from a pre-populated entity list rather than having to select a span manually. We empirically show that annotations performed in this manner are faster and more coherent with higher inter-annotator agreements. We benchmark the collected data on state-of-the-art models and release it in the open domain at https://github.com/rahular/model-based-coref.

Part III

OTHER TOPICS

6

FOCUS ATTENTION: PROMOTING FAITHFULNESS AND DIVERSITY IN SUMMARIZATION

6.1 ABSTRACT

Professional summaries are written with document-level information, such as the theme of the document, in mind. This is in contrast with most seq2seq decoders which simultaneously learn to focus on salient content, while deciding what to generate, at each decoding step. With the motivation to narrow this gap, we introduce Focus Attention Mechanism, a simple yet effective method to encourage decoders to proactively generate tokens that are similar or topical to the input document. Further, we propose a Focus Sampling method to enable generation of diverse summaries, an area currently understudied in summarization. When evaluated on the BBC extreme summarization task, two state-of-the-art models augmented with Focus Attention generate summaries that are closer to the target and more faithful to their input documents, outperforming their vanilla counterparts on ROUGE and multiple faithfulness measures. We also empirically demonstrate that Focus Sampling is more effective in generating diverse and faithful summaries than top-k or nucleus sampling-based decoding methods.

6.2 INTRODUCTION

Document summarization — producing the shorter version of a document while preserving salient information (Mani, 2001; Nenkova and McKeown, 2011) — is challenging even for humans. Today, systems can generate summaries with a high level of fluency and coherence. This is due to recent advances such as sequence-to-sequence architectures (seq2seq) with attention and copy mechanism (Bahdanau et al., 2015; Gu et al., 2016; Hochreiter and Schmidhuber, 1997a), fully attention-based Transformer architectures (Vaswani et al., 2017), and large pretrained language models (Devlin et al., 2018; Dong et al., 2019a; Lewis et al., 2019; Liu et al., 2019; Radford et al., 2018; Raffel et al., 2019a; Rothe et al., 2020; Song et al., 2019; Yang et al., 2019; Zhang et al., 2019c).

However, in terms of summary quality, many challenges remain. For example, generating summaries that are faithful to the input is an unsolved problem (Gabriel et al., 2020; Kryscinski et al., 2020; Maynez et al., 2020). Furthermore, there can be multiple equally good summaries per source document. Neural generation models fail to account for this and tend to generate outputs with low diversity due to standard likelihood training, approximate decoding objectives, and lack of high quality multi-reference datasets (Choi et al., 2020; Fan et

A	**Gold**: Australia has expelled an Israeli diplomat saying Israel was behind the forging of Australian passports linked to the murder of a Hamas operative in Dubai. **Pegasus**: Australia has expelled an Israeli diplomat after concluding that forged Australian passports used in the killing of a Hamas militant in Dubai were issued by Israel. **Our PegFame model**: The Australian government has expelled an Israeli diplomat over the use of forged Australian passports in the killing of a Hamas militant in Dubai.
B	**Pegasus with Top-k Sampling** Israel has summoned the Australian ambassador to complain after the Australian government said forged passports used in the killing of a Hamas operative in Dubai belonged to Netanyahu's foreign ministry. The Australian government has ordered Israel to withdraw an officer over the use of forged Australian passports used by the 2013 murder of a Lebanese opposition figure in Dubai. **Pegasus with Nucleus Sampling** Israel hasracuse withdrawn an envoy after the Australian government said it concluded that Israeli agents used forged passports used to kill a Dubai Bendigo businessman. The Australian government has recalled an Israeli diplomat over accusation that fake Australian passports used 436 kilometres (300 miles) from Canberra in the death of a Hamas militant were stolen by Israeli agents.
C	**Our PegFame model with novel Focus Sampling** Australia has expelled an Israeli diplomatic staff after accusing the country's security agency, the Israeli military's intelligence agency, of being responsible for the use of Australian visas used in the killing of a Palestinian. The Australian government has expelled an Israeli diplomatic staff after it said the country was responsible for the use of Australian visas used in the killing of a Palestinian in the Middle East.

Figure 18: Block A shows the best predictions from PEGASUS and our PEGFAME (PEGASUS with FAME) model, along with the GOLD summary for an XSUM article. Block B presents diverse summaries generated from PEGASUS using top-k and nucleus sampling. Block C shows diverse summaries generated using our PEGFAME model with Focus sampling. The text in orange is not supported by the input article.

al., 2018; Freitag et al., 2020; Kulikov et al., 2019). Not much attention has been given to generation of diverse, yet faithful summaries – two goals are often challenging to achieve simultaneously (Hashimoto et al., 2019); a model can produce diverse outputs through sampling (Fan et al., 2018; Holtzman et al., 2020), but at the cost of quality.

In this paper we introduce a Focus Attention MEchanism (or FAME) to transformer-based seq2seq architectures. FAME is inspired by how humans write summaries. Specifically, FAME aims to perform source-side planning to focus the summary on supported and topical content. FAME achieves this through a novel technique which augments standard contextual representations with a dynamic source-conditioned vocabulary biasing layer. We present the following experimental findings:

FAME PROMOTES SUMMARIES FAITHFUL TO THE SOURCE When evaluated on the BBC extreme summarization task (XSUM; Narayan et al., 2018), experiments with two state-of-the-art summarizers – ROBERTAS2S (Rothe et al., 2020) and PEGASUS (Zhang et al., 2019c) – show that both models generate summaries that are more faithful to their

input documents when augmented with FAME, in comparison with their vanilla counterparts.[1] Faithfulness is measured through a variety of previously proposed metrics. In addition, we leverage the manually annotated document-summary pairs for faithfulness from Maynez et al. (2020) and train a scorer which serves as an efficient proxy for expensive human evaluations. We call this metric *BERT-Faithful*.

FAME ENABLES DIVERSE SUMMARIES FAME, by design, supports *Focus Sampling* – a technique that is more effective in sampling topically relevant tokens to generate diverse, yet topically consistent and faithful outputs, than other sampling methods (Fan et al., 2018; Holtzman et al., 2020). Figure 18 illustrates how focus sampling generates better summaries than other sampling methods. We demonstrate the effectiveness of our new Focus Sampling technique using a variety of existing diversity and faithfulness measures. Empirically, we find that optimizing for high diversity often comes at the cost of faithfulness. Thus FAME provides a mechanism for trading-off high faithfulness with better diversity in summarization.

6.3 RELATED WORK

6.3.1 *Task-Specific Architectural Priors*

Several works enhance seq2seq architectures with task-specific priors. Pointer-generator style models (See et al., 2017; Xu et al., 2020) can accurately generate mostly extractive summaries by copying words from the source text via pointing. Text editing models (Dong et al., 2019b; Mallinson et al., 2020; Malmi et al., 2019) cast text generation as a sequence tagging problem with carefully selected edit operations required for the task. Others focus on improving content selection to better constrain the model to likely input phrases (Gehrmann et al., 2018) or by improving the representation of relevant input tokens (Zhou et al., 2017). Instead of directly modeling such priors, FAME learns the theme of the document through dynamic vocabulary biasing. Thus, FAME can be seen as a generalization of Pointer-generator or text-editing models via soft vocabulary learning. In fact, our FAME models achieve state-of-the-art on text-editing tasks (Appendix A.3.3).

6.3.2 *Topic-Aware Generation Models*

The idea of capturing document-level semantic information has been widely explored in the summarization community. Barzilay and Elhadad (1997) use WordNet (Fellbaum, 1998) to model a text's content relative to a topic based on lexical chains. Lin and Hovy (2000) propose to learn topic signatures for summarizing documents. Recently,

1 In the paper we focus on assessing FAME on XSUM. But other summarization and text editing results can be found in Appendix A.3.2 and A.3.3.

document-level topic information has been used for improving neural language models (Dieng et al., 2017; Ghosh et al., 2016; Karmaker Santu et al., 2019; Mikolov and Zweig, 2012), neural response generators (Dziri et al., 2019; Xing et al., 2017), and not surprisingly, neural summarizers (Ailem et al., 2019; Narayan et al., 2018; Wang et al., 2020e). Both, Narayan et al. (2018) and Ailem et al. (2019), use a pretrained Latent Dirichlet Allocation (LDA; Blei et al., 2003) model, whereas, Wang et al. (2020e) use Poisson factor analysis (Zhou et al., 2012), to synthesize topic vectors for the input. Instead, we dynamically learn a target-induced topic distribution for the input under the assumption that the human-written summary is a good proxy for the input document.

6.3.3 *Faithful Generation Models*

Cao et al. (2017) force faithful generation by conditioning on both source text and extracted fact descriptions from the source text. Song et al. (2020) propose to jointly generate a sentence and its syntactic dependency parse to induce grammaticality and faithfulness. Tian et al. (2019) learn a confidence score to ensure that the model attends to the source whenever necessary. Wang et al. (2020f) introduce new input-output matching and embedding similarity losses to alleviate hallucination issues. Yet, the task of generating text that is consistent with the input remains an open problem (Gabriel et al., 2020).

6.3.4 *Diverse Generation Models*

There has been a surge of interest in making language models generate more diverse and human-like outputs. Vijayakumar et al. (2018) and Kulikov et al. (2019) diversify beam search, using a task-specific scoring function, or constrain beam hypotheses to be sufficiently different. Others avoid text degeneration by truncating the unreliable tail of the probability distribution at each decoding step, either by sampling from the top-k tokens (*Top-k Sampling*; Fan et al., 2018) or by sampling from a dynamic nucleus of tokens with the bulk of the probability mass (*Nucleus Sampling*; Holtzman et al., 2020). Others modify the training objective to make the distribution sparse (Martins et al., 2020) or assign lower probability to unlikely generations (Welleck et al., 2019a).

For conditional text generation, most work focuses on generating diverse questions (Dong et al., 2017; Narayan et al., 2016; Sultan et al., 2020; Wang et al., 2020d) or paraphrases (Cao and Wan, 2020; Dai et al., 2017; Li et al., 2016b; Xu et al., 2018). Following Gehrmann et al. (2018), Cho et al. (2019) use a mixture of experts to sample different binary masks on the source sequence for diverse content selection for summarization.

Our focus sampling is similar to top-k and nucleus sampling methods; in that it truncates the tail of the probability distribution. However, instead of truncating it at each decoding step, it biases the de-

coder proactively to generate output from a set of tokens which are topically-relevant to the input.

6.4 SUMMARIZATION WITH FOCUS ATTENTION

Given an input document $X_{1:n}$, we aim to generate its summary $Y_{1:m}$, where n and m are input and output sequence lengths. We address this problem using seq2seq architectures with Transformer encoder and decoder, augmented with FAME, as depicted in Figure 19. FAME learns a distribution $\mathbf{t}_{x_i}$ for each input token x_i over the vocabulary, measuring similarity of x_i (in context) to the tokens in the vocabulary. The vocabulary distributions, $\mathbf{t}_{x_i}$, for all x_i are combined to form a dynamic vocabulary bias that is added to the decoder logits. This mechanism enhances the conditioning on the input source and encourages the decoder to generate tokens that are topically similar to the input.

6.4.1 *Transformer-based seq2seq Model*

The encoder uses BERT Transformer layers with multi-headed self-attention to encode X to a vector sequence $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_n$, with $\mathbf{x}_i \in \mathbb{R}^h$, where h is the size of hidden representation. The decoder uses an identical architecture, except that at decoding step t, layer l adds a conditional representation $\mathbf{y}_t^l \in \mathbb{R}^h$ for the token y_t by attending to the output representation $\mathbf{Y}_{1:t-1}^{l-1} = \mathbf{y}_1^{l-1}, \ldots, \mathbf{y}_{t-1}^{l-1}$ generated so far through self-attention and by attending to the input contextual representation $\mathbf{X}$ through encoder-decoder attention. The probability of predicting the next token y_t from a vocabulary V is:

$$p(y_t|Y_{1:t-1}, X; \theta) = \text{softmax}(\mathbf{E}\mathbf{y}_t^L), \tag{6.3}$$

where, $\mathbf{y}_t^L$ is the representation from the final decoder layer L, $\mathbf{E} \in \mathbb{R}^{|V| \times h}$ the embedding matrix and θ the model parameters. Parameters are trained by minimizing cross-entropy at each decoding step:

$$L_{\text{MLE}}(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \log p(\hat{y}_t|\hat{Y}_{1:t-1}, X; \theta), \tag{6.4}$$

where, $\hat{Y}_{1:m}$ is the human-written summary.

6.4.2 *Focus Attention MEchansim (*FAME*)*

It is challenging for a decoder to obtain all relevant information from the conditional representation $\mathbf{y}_t^L$ to learn the vocabulary output logits such that predictions y_t are consistent with the input. Other modeling factors, specifically the decoder language model, can overwhelm model predictions. FAME (Figure 19) addresses this by introducing a short-circuit from the source to the vocabulary output logits via a source-conditioned bias on vocabulary items.

We take the encoder representation $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_n$ and learn a *Token-level Vocabulary Distribution* $\mathbf{t}_{x_i} = \text{gelu}(\mathbf{x}_i \mathbf{W}_1)\mathbf{W}_2\mathbf{E} \in \mathbb{R}^{|V|}$, for

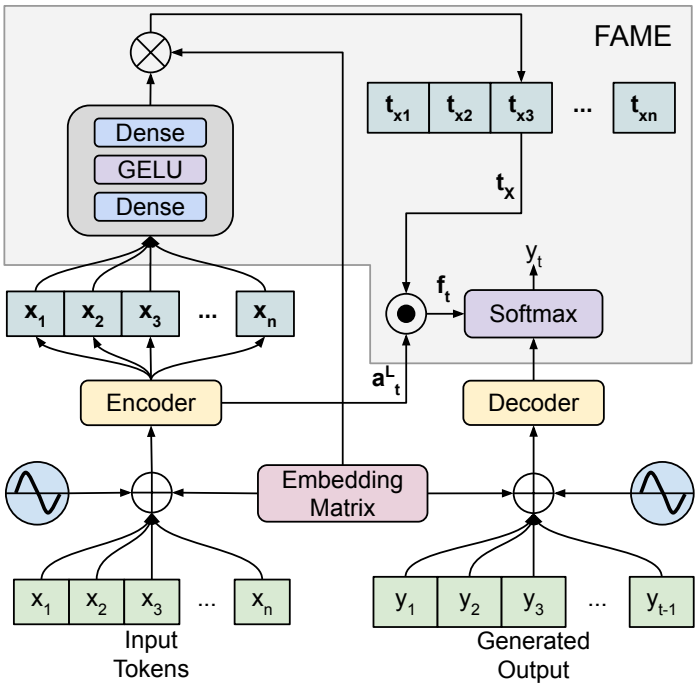


Figure 19: A Transformer-based encoder-decoder architecture with FAME.

each token x_i in the input sequence $\mathbf{X}$. $\mathbf{t}_{x_i}$ measures the contextual similarity of the input token x_i to the tokens in the vocabulary; $\boldsymbol{W}_1 \in \mathbb{R}^{h \times h'}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{h' \times h}$ are parameters of newly introduced dense layers, h' is the intermediate filter size. We define a *Source-conditioned Vocabulary Distribution* as $\mathbf{t}_X = 1/n \sum_{i=1}^{n} \mathbf{t}_{x_i} \in \mathbb{R}^{|V|}$ as an average of token-level vocabulary distributions for tokens present in the input sequence X, capturing the similarity of X to the tokens in the vocabulary.

Let $\mathbf{a}_t^L \in \mathbb{R}^n$ be the encoder-decoder attention distribution over the source tokens for the output token y_t and the final decoder layer L. We use $\mathbf{a}_t^L$ to produce a weighted sum of the token-level vocabulary distributions to compute a dynamic vocabulary bias, or *Focus Bias* $\mathbf{f}_t = \sum_{i=1}^{n} \mathbf{a}_{t,i}^L \mathbf{t}_{x_i} \in \mathbb{R}^{|V|}$ at decoding step t. We modify the probability of predicting the next token y_t from a vocabulary V as:

$$p(y_t | Y_{1:t-1}, X; \theta) = \text{softmax}(\mathbf{y}_t^L \mathbf{E} + \mathbf{f}_t) \tag{6.5}$$

We call this *Focused Probability Distribution*, and it modifies the output logits dynamically to put more focus on those tokens in the vocabulary which are similar to the attended tokens in X. The focus bias introduces a human-inspired control to the model where we do not generate the output in a fully abstractive manner (as in equation 6.3), but we proactively generate output tokens that are similar to the input tokens (as in equation 6.5).

6.4.3 *Summary-induced Topic Focused Distribution*

We aim to guide our focus bias $\mathbf{f}_t$ to be a better representative of the topical content relevant for the task. We achieve this by using

the human-written summary $\hat{Y}$ as a proxy for the topical content of the input and impose the following prior on the source-conditioned vocabulary distribution $\mathbf{t}_X$:

$$L_{\text{Topic}}(\theta) = -\frac{1}{|V|}\sum_{i=1}^{|V|}([v_i \in \hat{Y}]\log(\sigma(\mathbf{t}_{X,i})) + [v_i \notin \hat{Y}]\log(1-\sigma(\mathbf{t}_{X,i}))). \tag{6.6}$$

We further refine equation 6.6 by replacing $\hat{Y}$ with $\hat{Y}_c = \hat{Y} - F$, where F is a set of $|F|$ most frequent tokens in the vocabulary,[2] to improve focus on content words. Our final loss function is then

$$L = \lambda L_{\text{MLE}} + (1-\lambda)L_{\text{Topic}}, \tag{6.7}$$

where, λ is an hyper parameter.[3]

By enforcing $\mathbf{t}_X$ to be a topic distribution for the input X, we encourage the focus bias $\mathbf{f}_t$ to promote topically relevant tokens, and subsequently generate topically consistent outputs. Importantly, our focus bias with target-induced topic distribution is task-agnostic and less vulnerable to reference divergence issues (Dhingra et al., 2019; Maynez et al., 2020), and can learn any property embodied in the target relevant for the task. For example, depending on the task, $\mathbf{f}_t$ can learn to favour input tokens (e.g., for mostly extractive summaries) or new tokens (e.g., for mostly abstractive summaries). This is in sharp contrast to models that introduce task-specific priors, e.g., the pointer-generator network (See et al., 2017) that can copy words from the source text, but does not do well on extreme summarization which is highly abstractive in nature (Narayan et al., 2018).

6.4.4 *Focus Sampling: Promoting Diversity in Faithful Generation*

We introduce *Focus Sampling* with FAME to construct a subset $V_k \subseteq V$ by sampling k tokens from the topic distribution $\mathbf{t}_X$ ($\text{Focus}_{\text{sample},k}$). Then, we modify equation 6.5 as

$$p(y_t|Y_{1:t-1}, X; \theta) = \begin{cases} \text{softmax}(\mathbf{y}_t^L\mathbf{E} + f_t)_i & \text{if } v_i \in V_k \cup F \\ 0, & \text{otherwise.} \end{cases} \tag{6.8}$$

For document summarization, the subset V_k will capture topically salient tokens necessary to generate a summary; F is always added to V_k to ensure that the model has access to function words. By tuning the parameters of sampling, we can enforce the model to control the faithfulness or diversity of the outputs.

Focus sampling has similarities to top-k ($\text{Div}_{\text{top},k}$; Fan et al., 2018) and nucleus sampling ($\text{Div}_{\text{nucleus}}$; Holtzman et al., 2020); in that they all aim to promote diversity. At each decoding step, the top-k sampling diversifies the generation process by sampling a token from the

2 which are usually articles or other function words.

3 λ is set to 0.5 for all experiments.

top k tokens in the final output distribution. Similarly, nucleus sampling samples from a dynamic nucleus of tokens containing the vast majority (with a cumulative probability p) of the probability distribution. Both top-k and nucleus sampling shorten the tail of the output distribution at each decoding step, whereas focus sampling constrains the decoder to use a fixed and topically relevant vocabulary V_k. Unlike the other two techniques, $\text{Focus}_{\text{sample,k}}$ can also benefit from standard beam search decoding, leading to superior generation that is not only diverse, but also consistent with the input document.

6.5 EXPERIMENTAL SETUP

In this section we present our experimental setup to assess the ability of our FAME models to generate faithful summaries and to demonstrate that focus sampling is more effective in generating diverse and faithful summaries than other sampling-based decoding methods.

6.5.1 *Extreme Summarization*

We evaluate FAME models on extreme document summarization (XSUM; Narayan et al., 2018). The XSUM summaries, are extreme in that the documents are summarized into single-sentence summaries. These summaries demonstrate a high level of abstractiveness, and generating them automatically requires document-level inference, abstraction, and paraphrasing. Due to their extreme nature, XSUM summaries are ideal to evaluate FAME models' ability to capture the theme of the document.[4] We use on the original cased version consisting of 204,045/11,332/11,334 training/validation/test document-summary pairs. During training, the input documents are truncated to 512 tokens. The length of the summaries are limited to 64.

6.5.2 *Pretrained Models with* FAME

We introduce FAME to two popular seq2seq architectures: RoBERTa initialized seq2seq (RobertaS2S, Rothe et al., 2020) and PEGASUS (Zhang et al., 2019c). We refer RobertaS2S models with FAME as RobFame and PEGASUS with FAME with PegFame.

We experiment with RobertaS2S-Large with shared encoder and decoder; it has 24 layers, a hidden size of 1024, filter size of 4096, 16 attention heads, and a vocabulary with 50K sentence pieces (Kudo and Richardson, 2018). RobertaS2S has around 455M parameters and RobFame has an additional 8M parameters.

The best-performing PEGASUS model from Zhang et al. (2019c) is not directly comparable with RobertaS2S. It does not share the encoder and decoder, it only has 16 layers, a hidden size of 1024, filter

4 We further experiment with long-form story highlight generation (CNN/DM; Hermann et al., 2015) and two text editing tasks: Sentence Fusion (Geva et al., 2019) and Sentence Splitting (Botha et al., 2018). Their results can be found in Appendix A.3.2 and A.3.3. Our FAME models achieve SOTA on both text-editing tasks.

size of 4096, 16 attention heads, with a total of 568M parameters, and it also uses a much larger vocabulary with 91K sentence pieces. Hence, we trained our own PEGASUS model. We use the same architecture as ROBERTAS2S and pretrain it on a mixture of C4 (Raffel et al., 2019a) and HugeNews (Zhang et al., 2019c) datasets with the original objective of generating salient GAP-sentences.

Our experiments focus on this newly trained PEGASUS model which has same number of parameters and vocabulary as ROBERTAS2S. But in contrast to ROBERTAS2S, the encoder-decoder attention in PEGASUS is pretrained. This allows us to analyse how focus attention affects pretrained (PEGASUS) vs randomly-initialized (ROBERTAS2S) encoder-decoder attentions.[5]

6.5.3 *Evaluation Metrics*

LEXICAL OVERLAP We report ROUGE F1 scores (Lin and Hovy, 2003) against reference summaries; in particular, we report on ROUGE-1 and ROUGE-2 for informativeness and ROUGE-L for fluency.[6]

SEMANTIC SIMILARITY We report *BERTScore* (Zhang et al., 2020b) which computes the contextual similarity between a candidate and its reference summary.

FAITHFULNESS ROUGE and BERTScore do not correlate well with faithfulness of the generated summaries (Maynez et al., 2020). Human evaluation is traditionally considered as the gold standard for measuring faithfulness. But recent research has shown that even human evaluation has shortcomings (Schoch et al., 2020). Moreover, it is prohibitively expensive. This has led to the proposal of meta-evaluation metrics for various generation tasks (Durmus et al., 2020; Kryściński et al., 2019; Rei et al., 2020; Sellam et al., 2020).

We evaluate FAME models on semantic inference metrics such as textual entailment (Falke et al., 2019; Kryscinski et al., 2019; Pasunuru and Bansal, 2018; Welleck et al., 2019b) and question answering (Arumae and Liu, 2019; Wang et al., 2020a). In particular, we report the probability of a summary entailing (*ent.*) its input document (Maynez et al., 2020) and QA-based *Feqa* scores (Durmus et al., 2020). For ent. scores, we train an entailment classifier by fine-tuning a BERT-Large pretrained model (Devlin et al., 2018) on the Multi-NLI dataset (Williams et al., 2018). For Feqa, we use a fine-tuned BART (Lewis et al., 2019) language model for question generation to generate questions from the summaries, and a BERT-base model fine-tuned on SQuAD (Rajpurkar et al., 2016b) to answer the generated questions with input document as context.[7]

In addition to *ent.* and *Feqa*, we train a scorer leveraging manually annotated document-summary pairs for faithfulness, as a surrogate

5 See Appendix A.3.1 for implementation details and hyperparameter settings.

6 We lowercased candidate and reference summaries and used `pyrouge` with parameters "-a -c 95 -m -n 4 -w 1.2."

7 We used the Feqa code available here: https://github.com/esdurmus/feqa/.

for human evaluation and call this metric *BERTFaithful*.[8] In particular, we finetune a BERT-Base classifier on 500 manually annotated document and gold summary pairs for the XSum dataset from Maynez et al. (2020) to predict whether a summary is faithful to the input document or not.[9] We report the percentage of summaries that were faithful ($\frac{1}{N}\sum_i \mathbb{1}[p_i(\text{faithful}) > 0.5]$) and the model's confidence to generate faithful summaries ($\frac{1}{N}\sum_i p_i(\text{faithful})$); N is the total number of examples in the test set.

DIVERSITY We report the number of times (out of n), a model is able to generate a completely new summary (*Unique*), and *Distinct-N* (Li et al., 2016a), measuring the lexical diversity in the generated summaries. Distinct-N is estimated as the number of distinct n-grams of order n divided by the total number of n-grams of the same order, in all generated summaries.

Finally, we also report the average length of summaries (*Len.*), repetition errors (*Rep.*, estimated as the percentage of summaries with at least one repetition of rare or content words), and ROUGE-1 precision against the input document (*R1, P%*), to better understand their quality.

6.6 RESULTS

Models	Lexical			Sem.	Faithfulness				others		
	Overlap (w/ ref)			Sim.	ent.	Feqa	BERT-F		Len.	Rep.	R1
	R1	R2	RL	BERTSc.			%	conf.		(↓)	(P%)
RobertaS2S	41.45	18.79	33.90	80.6	39.1	19.8	21.5	0.216	21.2	24.2	71.1
RobFame	42.15	19.68	34.81	80.8	41.3	21.2	22.7	0.226	20.8	20.7	72.5
Pegasus	44.85	22.26	37.03	81.7	43.6	24.5	27.0	0.263	21.1	6.0	73.8
PegFame	**45.31**	**22.75**	**37.46**	**81.9**	**44.8**	**24.8**	**27.3**	**0.269**	20.8	**5.3**	**74.3**

Table 11: Abstractive Summarization results on XSUM test set comparing FAME models with their baselines. For all our models, we use standard beam decoding with a beam size of 4 to generate the single best summary for a document. Focus sampling is not used here. See Section 6.5.3 for details on the evaluation metrics reported. Best number for each metric is **boldfaced**. (BERTSc. and BERT-F stand for BertScore and BERTFaithful respectively.)

8 A very similar scorer was used in the GEM benchmark (Gehrmann et al., 2021) to identify and extract the subset with faithful reference summaries from the XSum dataset (Narayan et al., 2018).

9 Out of 500, 90% of the document-summary pairs were used for training and the rest 50 document-summary pairs were used for validation. We used the validation set to estimate Spearman's correlation coefficients of different metrics with the human assessment for faithfulness. We found that both entailment scores (*ent.*) and *BERT-Faithful* are moderately correlated with faithfulness with correlation coefficients of 0.4387 and 0.3889, respectively. As such, we believe that BERTFaithful works as an efficient proxy for expensive human evaluation for faithfulness for XSum summaries. More work is needed to understand if BERTFaithful generalizes to other datasets.

Metrics	Uni.	Dist.-N			ROUGE			ent.	B-Sc.
		1	2	3	R1	R2	RL		
RobertaS2S									
($Div_{top,k}$)	9.98	2.5	25.0	57.7	33.6	12.0	26.5	21.8	76.9
($Div_{nucleus}$)	9.99	**4.1**	30.1	62.2	32.4	11.4	25.6	19.7	75.7
RobFame									
($Div_{top,k}$)	9.99	2.3	25.0	58.1	32.7	11.3	25.7	20.3	76.6
($Div_{nucleus}$)	9.99	**4.1**	**30.7**	**63.2**	31.3	10.6	24.7	18.0	75.4
($Focus_{sample,k}$)	1.61	3.5	22.4	43.9	**38.0**	**15.7**	**31.0**	**34.3**	**78.6**
($Focus_{sample,k}$, $Div_{top,k}$)	9.99	2.1	20.3	51.8	31.8	10.2	24.7	24.3	75.4
($Focus_{sample,k}$, $Div_{nucleus}$)	9.98	1.9	18.4	48.2	32.9	11.1	25.8	25.9	76.1
Pegasus									
($Div_{top,k}$)	9.98	1.9	23.2	55.3	36.6	14.3	28.8	27.7	78.4
($Div_{nucleus}$)	9.99	**3.8**	**30.5**	**63.1**	34.1	12.8	26.9	22.7	76.5
PegFame									
($Div_{top,k}$)	9.98	1.9	23.2	55.5	36.7	14.5	29.0	28.5	**78.5**
($Div_{nucleus}$)	9.99	**3.8**	30.4	**63.1**	34.2	12.8	27.0	23.2	76.6
($Focus_{sample,k}$)	2.77	2.4	16.5	34.2	**37.5**	**15.4**	**30.3**	**33.6**	77.9
($Focus_{sample,k}$, $Div_{top,k}$)	8.99	2.8	23.0	54.7	31.5	10.3	24.4	22.8	74.7
($Focus_{sample,k}$, $Div_{nucleus}$)	9.98	2.6	20.8	50.9	32.5	11.0	25.3	24.8	75.3

Table 12: Assessment of diversity, relevance and faithfulness with focus sampling on the XSUM test set. (Uni., ent., and B-Sc. represent unique summaries, entailment scores, and BERTScores respectively.)

FAME SUMMARIES ARE MORE FLUENT, INFORMATIVE AND FAITHFUL. Table 11 presents results comparing our FAME models, RobFame and PegFame, against their counterparts RobertaS2S and Pegasus, respectively. Both FAME models clearly outperform their vanilla counterparts in terms of generating summaries that are more fluent (see RL and Rep.), more informative (see R1, R2 and BERTSc.) and more faithful (see ent., Feqa and BERTFaithful). Among all four models, PegFame summaries are most fluent, informative and faithful.

We further did pairwise comparisons for all measures in Table 11 and found that all differences are statistically significant except for BERTScore and faithfulness measures between Pegasus and PegFame.[10] These assessments demonstrate that FAME models aid both RobertaS2S and Pegasus in generating fluent, faithful and relevant summaries, but are more effective in RobertaS2S than in Pegasus for extreme summarization.

GENERATING DIVERSE AND FAITHFUL SUMMARIES WITH FOCUS SAMPLING. Table 12 presents results assessing focus sampling ($Focus_{sample,k}$), top-k sampling ($Div_{top,k}$) and nucleus sampling ($Div_{nucleus}$), for their abilities to generate diverse and faithful summaries. For $Focus_{sample,k}$,

10 All significance tests in this work are pairwise comparisons (one-way ANOVA with posthoc Tukey HSD tests; $p < 0.01$).

we choose $k = 10,000$. We follow Holtzman et al. (2020) and choose $k = 640$ and the nucleus probability $p = 0.95$, for $\text{Div}_{\text{top,k}}$ and $\text{Div}_{\text{nucleus}}$, respectively. For $\text{Focus}_{\text{sample,k}}$, we decode with a beam size of 4. We also report $\text{Focus}_{\text{sample,k}}$ with $\text{Div}_{\text{top,k}}$ and $\text{Div}_{\text{nucleus}}$ to assess if they can benefit one-another. In each setting we sample 10 summaries for each input document. For all metrics, we report the average over all 10 samples.[11]

Both $\text{Div}_{\text{top,k}}$ and $\text{Div}_{\text{nucleus}}$ almost always generate a new summary. In comparison $\text{Focus}_{\text{sample,k}}$ generates 1.61 and 2.77 unique summaries using RobFame and PegFame models, respectively. $\text{Div}_{\text{nucleus}}$ tends to generate the most distinct unigrams, bigrams, and trigrams. Interestingly, $\text{Focus}_{\text{sample,k}}$ summaries have a more diverse collection of unigrams than in $\text{Div}_{\text{top,k}}$ summaries (3.5% vs 2.3% for RobFame and 2.4% vs 1.9% for PegFame).

The high diversity in $\text{Div}_{\text{top,k}}$ and $\text{Div}_{\text{nucleus}}$ comes at the cost of faithfulness; summaries generated with these sampling techniques have poor entailment scores. $\text{Focus}_{\text{sample,k}}$, on the other hand, generates summaries which entail documents the most. It also has the highest ROUGE scores across the board. Some of the generated examples can be seen in Figure 18. More predictions from other models can be found in Appendix A.3.5. Augmenting $\text{Div}_{\text{top,k}}$ and $\text{Div}_{\text{nucleus}}$ with $\text{Focus}_{\text{sample,k}}$ is not desirable because, though it increases diversity in terms of uniqueness and Distinct-3 scores, faithfulness suffers again.

Comparing results in Table 12 to the results in Table 11, it is clear that diversity comes at the cost of quality (e.g., RL/ent. scores for RobFame and RobFame-$\text{Focus}_{\text{sample,k}}$ are 34.81/41.3 and 31.0/34.3, respectively). However, $\text{Focus}_{\text{sample,k}}$ is superior to both $\text{Div}_{\text{top,k}}$ and $\text{Div}_{\text{nucleus}}$ in generating better quality summaries.

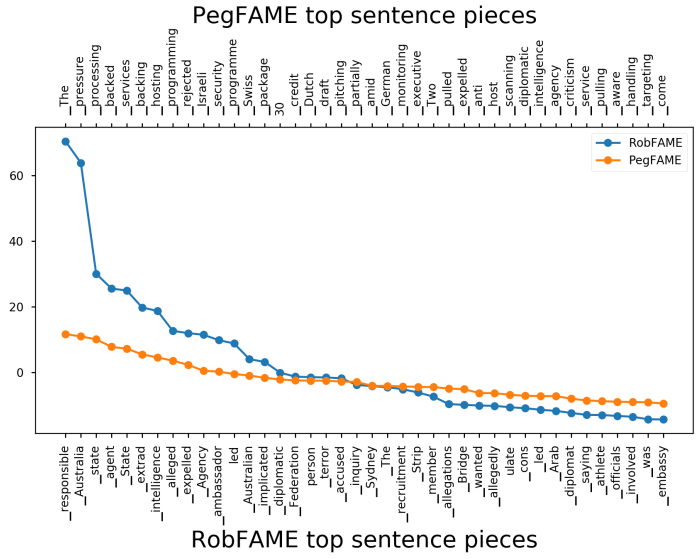


Figure 20: Top 40 sentence pieces and their logits from topic distribution t_X in RobFame and PegFame for the XSum article discussed in Figure 18.

11 Feqa and BERTFaithful scores are dropped due to time constraints.

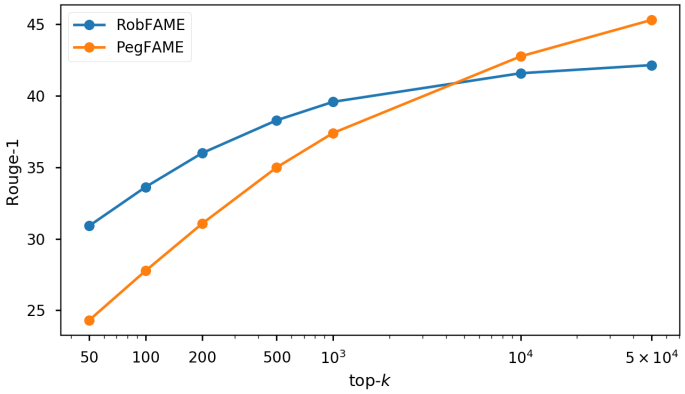


Figure 21: ROUGE-1 F1 scores of RobFame and PegFame models with different top-k vocabularies (equation 6.8) on the XSum test set. Similar patters are observed for ROUGE-2 and ROUGE-L scores.

FOCUS ATTENTION AND SAMPLING WORK DIFFERENTLY IN ROBFAME AND PEGFAME. Since both encoder-decoder and focus attention parameters of RobFame are randomly initialized, they learn to compliment each other and learn a peaky topic distribution. On the other hand, since PegFame's encoder-decoder attention is pre-trained, there is a push-pull effect between it and focus attention. This results in a smoother topic distribution, as seen in Figure 20.[12]

Although we see that both models' token sets capture the target intent well, the peaky distribution of RobFame enables more accurate predictions than that of PegFame, in a controlled generation setting. A comparison is presented in Figure 21 where we show how ROUGE-1 scores vary when we use only top-k tokens from t_X for generation.[13] We observe that RobFame consistently outperforms PegFame with the lower values of $k \in \{50, 100, 200, 500, 1000\}$.

Further, we observe that RobFame generates fewer unique summaries (1.61 vs 2.77) but has higher Distinct-N scores (3.5/22.4/43.9 vs 2.4/16.5/34.2) than PegFame, with $\text{Focus}_{\text{sample},k}$ in Table 12. This can be again be attributed to how Fame works differently in RobFame and PegFame. When V_k is sampled from RobFame's peaky distribution, the beam search decoding often tends to generate similar summaries (leading to a lower Uniqueness score) as the sampled V_ks do not diverge by much from each other. But when it does diverge, the decoder tends to generate completely new summaries (leading to higher Distinct-N scores).

Currently, we set $k = 10,000$ for our focus sampling experiments following our observations in Figure 21. Future work will focus on how to better leverage trade-off between diversity and faithfulness by controlling the peakiness of the topic distribution t_X.

12 This difference in topic distributions is consistent across the whole test set. We compute the peakiness score of a topic distribution as the slope of the line connecting logits of the top-1st token to the top-100th token. The average peakiness scores across the XSum testset for RobFame and PegFame are 1.25 (51°) and 0.45 (24.3°), respectively.

13 Additional results and model predictions for these experiments can be found in Appendix A.3.4.

Models	R1	R2	RL
Lead	16.30	1.61	11.95
PtGen (See et al., 2017)	29.70	9.21	23.24
ConvS2S (Narayan et al., 2018)	31.89	11.54	25.75
MMN (Kim et al., 2019)	32.00	12.10	26.00
MASS (Song et al., 2019)	39.75	17.24	31.95
BART (Lewis et al., 2019)	45.14	22.27	37.25
PEGASUS (Zhang et al., 2019c)	**47.21**	**24.56**	**39.25**
ROBERTAS2S (Rothe et al., 2020)	41.45	18.79	33.90
ROBFAME (w/o equation 6.6)	41.27	18.86	33.90
ROBFAME	42.15	19.68	34.81
ORACLE	**72.22**	**42.22**	**53.89**
PEGASUS (ours)	44.85	22.26	37.03
PEGFAME (w/o equation 6.6)	44.54	22.00	36.83
PEGFAME	45.31	22.75	37.46
ORACLE	**82.39**	**60.61**	**69.19**

Table 13: Ablations and SOTA comparisons on XSUM dataset. The **underlined bold** results are from the best performing models from literature and the **bold** results are the best performing FAME models.

ABLATIONS AND SOTA COMPARISONS We emphasize that FAME or focus sampling does not aim to improve on state-of-the-results in terms of ROUGE, but to generate more faithful or diverse summaries while maintaining their quality. For completeness, we compare our ROBFAME and PEGFAME models to their ablations and other state-of-the-art models on XSUM in Table 13.

We report ROUGE scores for FAME in the ideal scenario (ORACLE) where it focuses on all the correct tokens in the input, i.e., the topic distribution $\mathbf{t}_X$ is identical to the distribution observed in the reference summary. These models generate summaries with very high ROUGE scores when the model is given the correct tokens to focus on. The gap between the ORACLE and FAME scores suggests that there is still a lot of work to be done in this space. Focus attention without any topical supervision (models w/o equation 6.6) is not significantly better than the baselines. But ROBFAME and PEGFAME (trained with joint supervision in equation 6.7) significantly outperform ROBERTAS2S and PEGASUS, respectively.

Our best model PEGFAME performs better than PtGen (See et al., 2017), ConvS2S (Narayan et al., 2018), MMN (Kim et al., 2019), MASS (Song et al., 2019) and BART (Lewis et al., 2019), but worse when the original PEGASUS (Zhang et al., 2019c). This can be expected as the number of parameters in PEGFAME is far less than that in the original PEGASUS.

6.7 CONCLUSION

We introduced FAME, a new attention mechanism which dynamically biases the decoder to proactively generate tokens that are topically similar to the input. FAME enhances the faithfulness of existing state-of-the-art abstract summarization models while improving their overall ROUGE scores. Finally, our newly introduced focus sampling technique is a better alternative to top-k or nucleus sampling to generate diverse set of faithful summaries.

ETHICAL CONSIDERATIONS

The nature of text generation leads to multiple ethical considerations when applied to applications. The main failure mode is that the model can learn to mimic target properties in the training data that are not desirable.

FAITHFULNESS AND FACTUALITY Since models create new text, there is the danger that they may neither be faithful to the source material nor factual. This can be exacerbated when the data itself has highly abstractive targets, which require the model to generate words not seen in the source material during training. This often leads the model to generate content inconsistent with the source material (Gabriel et al., 2020; Kryscinski et al., 2020; Maynez et al., 2020).

TRUSTWORTHY DATA If the data itself is not trustworthy (comes from suspect or malicious sources) the model itself will naturally become untrustworthy as it will ultimately learn the language and topics of the training data. For instance, if the training data is about Obama birther conspiracies, and the model is asked to generate information about the early life of Obama, there is a risk that such false claims will be predicted by the model.

BIAS IN DATA Similarly, biases in the data around gender, race, etc., risk being propagated in the model predictions, which is common for most NLP tasks. This is especially true when the models are trained from non-contemporary data that do not represent current norms and practices (Blodgett et al., 2020).

The above considerations are non-malicious, in that the model is merely learning to behave as its underlying source material. If users of such models are not aware of these issues and do not account for them, e.g., with better data selection, evaluation, etc., then the generated text can be damaging.

Generation models can also be misused in malicious ways. These include generating fake news, spam, and other text meant to mislead large parts of the general population.

7

MINIMAX AND NEYMAN–PEARSON META-LEARNING FOR OUTLIER LANGUAGES

7.1 ABSTRACT

Model-agnostic meta-learning (MAML) has been recently put forth as a strategy to learn resource-poor languages in a sample-efficient fashion. Nevertheless, the properties of these languages are often *not* well represented by those available during training. Hence, we argue that the *i.i.d.* assumption ingrained in MAML makes it ill-suited for cross-lingual NLP. In fact, under a decision-theoretic framework, MAML can be interpreted as minimising the expected risk across training languages (with a uniform prior), which is known as Bayes criterion. To increase its robustness to outlier languages, we create two variants of MAML based on alternative criteria: Minimax MAML reduces the *maximum* risk across languages, while Neyman–Pearson MAML *constrains* the risk in each language to a maximum threshold. Both criteria constitute fully differentiable two-player games. In light of this, we propose a new adaptive optimiser solving for a local approximation to their Nash equilibrium. We evaluate both model variants on two popular NLP tasks, part-of-speech tagging and question answering. We report gains for their average and minimum performance across low-resource languages in zero- and few-shot settings, compared to joint multi-source transfer and vanilla MAML. The code for our experiments is available at https://github.com/rahular/robust-maml.

7.2 INTRODUCTION

Knowledge transfer is ubiquitous in machine learning because of the general scarcity of annotated data (Caruana, 1997; Pratt, 1993; Ruder, 2019, *inter alia*). A prominent example thereof is transfer from resource-rich languages to resource-poor languages (Ponti et al., 2019b; Ruder et al., 2019; Wu and Dredze, 2019). Recently, Model-Agnostic Meta-Learning (MAML; Finn et al., 2017) has come to the fore as a promising paradigm: it explicitly trains neural models that adapt to new languages quickly by extrapolating from just a few annotated data points (Gu et al., 2018; Li et al., 2020b; Nooralahzadeh et al., 2020; Wu et al., 2020a).

MAML usually rests on the simplifying assumption that the source 'tasks' and the target 'tasks' are independent and identically distributed (henceforth, *i.i.d.*). However, in practice most scenarios of cross-lingual transfer violate this assumption: training languages documented in mainstream datasets do not reflect the cross-lingual variation, as they belong to a clique of few families, geographical areas, and typological features (Bender, 2009; Joshi et al., 2020). Therefore, the majority

of the world's languages lies outside of such a clique. As training and evaluation languages differ in their joint distribution, they are not exchangeable (Orbanz, 2012; Ponti, 2021, ch. 6). Therefore, there is no formal guarantee that MAML generalises to the very languages whose need for transfer is most critical.

In this work, we interpret meta-learning within a decision-theoretic framework (Bickel and Doksum, 2015). MAML, we show, minimises the expected risk across languages found in the training distribution. Hence, it follows a so-called Bayes criterion. What if, instead, we formulated alternative criteria geared towards outlier languages? The first criterion we propose, Minimax MAML, is designed to be robust to worst-case-scenario out-of-distribution transfer: it minimises the *maximum* risk by learning an adversarial language distribution. The second criterion, Neyman–Pearson MAML, upper-bounds the risk for an arbitrary subset of languages via Lagrange multipliers, such that it does not exceed a predetermined threshold.

Crucially, both of these alternative criteria constitute competitive games between two players: one minimising the loss with respect to the neural parameters, the other maximising it with respect to the language distribution (Minimax MAML) or Lagrange multipliers (Neyman–Pearson MAML). Since an absolute Nash equilibrium may not exist for non-convex functions (Jin et al., 2020), such as neural networks, a common solution is to approximate local equilibria instead (Schäfer and Anandkumar, 2019). Therefore, we build on previously proposed optimisers (Balduzzi et al., 2018; Gemp and Mahadevan, 2018; Letcher et al., 2019) where players follow non-trivial strategies that take into account the opponent's predicted moves. In particular, we enhance them with first-order momentum and adaptive learning rate and apply them on our newly proposed criteria.

We run experiments on Universal Dependencies (Zeman et al., 2020) for part-of-speech (POS) tagging and TyDiQA (Clark et al., 2020) for question answering (QA). We perform knowledge transfer to 14 and 8 target languages, respectively, which belong to under-represented and often endangered families (such as Tupian from Southern America and Pama–Nyugan from Australia). We report modest but consistent gains for the average performance across languages in few-shot and zero-shot learning settings and mixed results for the minimum performance. In particular, Minimax and Neyman–Pearson MAML often surpass vanilla MAML and multi-source transfer baselines, which are currently considered state-of-the-art in these tasks (Clark et al., 2020; Ponti et al., 2021; Wu and Dredze, 2019).

7.3 SKEWED LANGUAGE DISTRIBUTIONS

Cross-lingual learning aims at transferring knowledge from resource-rich languages to resource-poor languages, to compensate for their deficiency of annotated data (Ponti et al., 2019a; Ruder et al., 2019; Tiedemann, 2015). The set of target languages ideally encompasses most of the world's languages. However, the source languages avail-

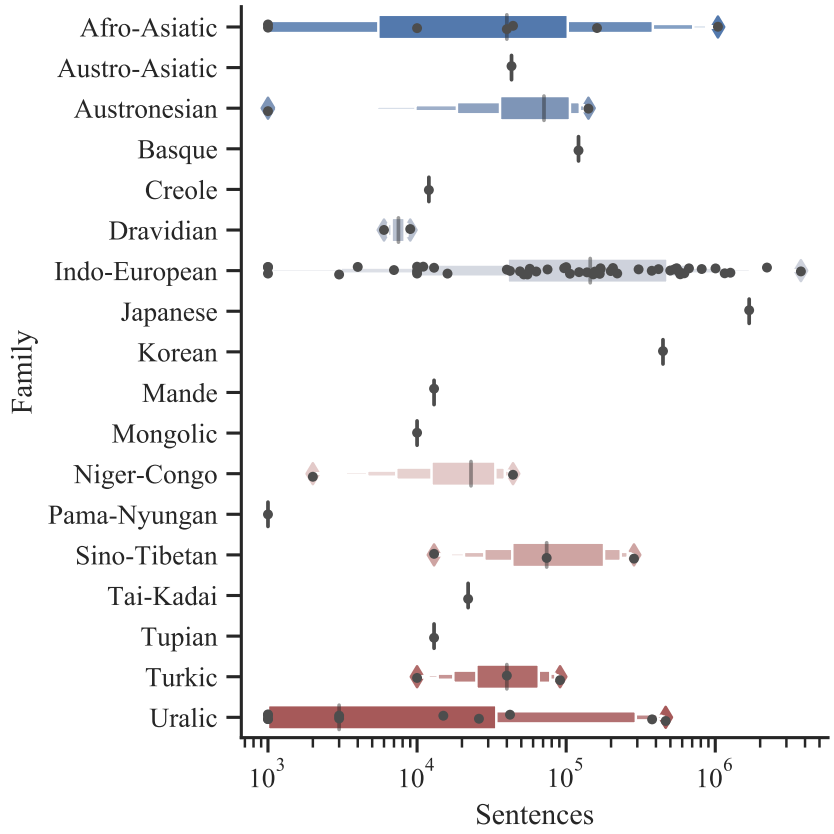


Figure 22: Annotated examples per family in the Universal Dependencies treebanks. Dots indicate individual languages, whereas boxes and whiskers mark quartiles.

able for training are often concentrated around few families, geographic areas, and typological features (Clark et al., 2020; Cotterell and Eisner, 2017; Gerz et al., 2018a,b; Ponti et al., 2020). As a consequence of this discrepancy, a language drawn at random might have no related languages available for training. Even when this is not the case, they might provide a scarce amount of examples for supervision.

To illustrate this point, consider Universal Dependencies (UD; Zeman et al., 2020), hitherto the most comprehensive collection of manually curated multilingual data. First, out of 245 families attested in the world according to Glottolog (Hammarström et al., 2016), UD covers only 18.[1] In fact, some families are chronically over-represented (e.g. Indo-European and Uralic) and others are neglected (e.g. Pama-Nyugan and Uto-Aztecan). Second, as shown in figure 22, the allocation of labelled examples across families is imbalanced (e.g. note the low counts for Niger–Congo or Dravidian languages). Third, one can measure how representative the linguistic traits of training languages are in comparison to those encountered around the globe. In figure 23, we represent UD languages as dots in the space of possible typological features in WALS (Dryer and Haspelmath, 2013). These are plotted against the density of the distribution based on all lan-

1 For more details on family distributions, cf. figure 36 in the Appendix.

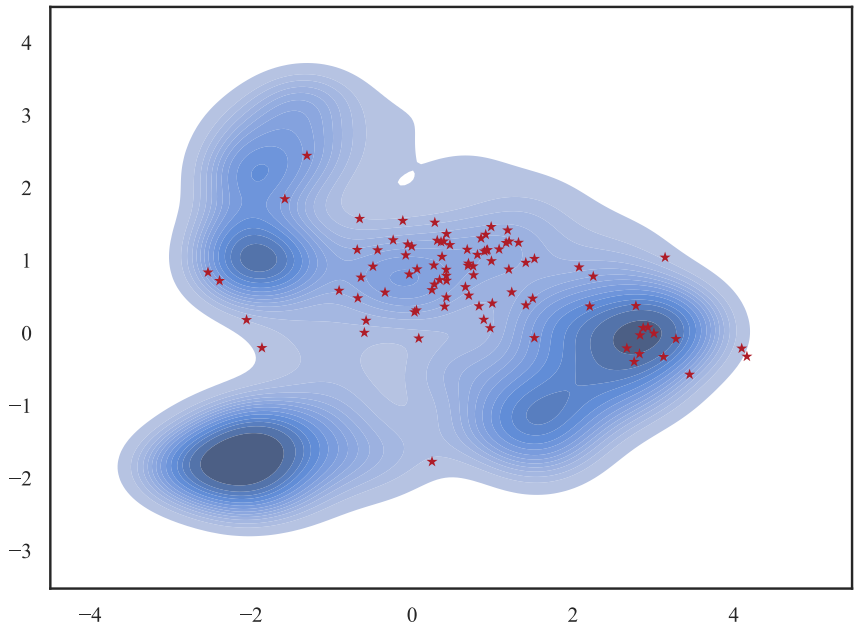


Figure 23: Density of WALS typological features of the world's languages reduced to 2 dimensions via PCA. Red dots are languages covered by UD. Darkness corresponds to more probable regions.

guages in existence. Crucially, it emerges that UD languages mostly lie in a low-density region. Therefore, they hardly reflect the variety of possible combinations of typological features.

In general, this demonstrates that the distribution of training languages in existing NLP datasets is heavily skewed compared to the real-world distribution. Indeed, this very argument holds true *a fortiori* in smaller, less diverse datasets. While this fact is undisputed in the literature, its consequences for modelling, which we expound in the next section, are often under-estimated.

7.4 ROBUST MAML

Model-Agnostic Meta Learning (MAML; Finn et al., 2017) has recently emerged as an effective approach to cross-lingual transfer (Gu et al., 2018; Li et al., 2020b; Nooralahzadeh et al., 2020; Wu et al., 2020a). MAML seeks a good initialisation point for neural weights in order to adapt them to new languages with only a few examples. To this end, for each language $\mathcal{T}_i$ a neural model f_θ is updated according to the loss on a batch of examples $\mathcal{L}_{\mathcal{T}_i}(f_\theta, \mathcal{D}_{train})$. This inner loop is iterated for k steps. Afterwards, the loss incurred by the model on a held-out batch $\mathcal{D}_{val}$ is compounded with those of the other languages as part of an outer loop, as shown in equation (7.9):

$$\begin{aligned} \theta^\star &= \min_\theta \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}(f_{\varphi_i}, \mathcal{D}_{val})\, p(\mathcal{T}_i) \\ \text{where} \quad \varphi_i &= \theta - \eta \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta, \mathcal{D}_{train}) \end{aligned} \tag{7.9}$$

where $\eta \in \mathbb{R}_{>0}$ is the learning rate. Language probabilities are often taken to follow a discrete uniform distribution $p(\mathcal{T}_i) = \frac{1}{|\mathcal{T}|}$. In this case, equation (7.9) becomes a simple average.

MAML can also be interpreted as point estimate inference in a hierarchical Bayesian graphical model (see figure 35 in the Appendix). In this case, the adapted parameters $\boldsymbol{\varphi}_i$ are equivalent to an intermediate language-specific variable acting as a bridge between the language-agnostic parameters θ and the data (Finn et al., 2018; Grant et al., 2018; Yoon et al., 2018). This allows us to reason about the conditions under which a model is expected to generalise to new languages. Crucially, generalisation rests on the assumption of independence and identical distribution among the examples (including both train and evaluation), which is known as exchangeability (Zabell, 2005). However, as seen in section 7.3, most of the world's languages are outliers with respect to the training language distribution. Therefore, there is no solid guarantee that meta-learning may fulfil its purpose, i.e. generalise to *held-out* languages.

7.4.1 *Decision-Theoretic Perspective*

To remedy the mismatch between assumptions and realistic conditions, in this work we propose objectives which can serve as alternatives to equation (7.9) of vanilla MAML. These are rooted in an interpretation of MAML within a decision-theoretic perspective (Bickel and Doksum, 2015, ch. 1.3), which we outline in what follows. The quantity of interest we aim at learning is the neural parameters θ. Therefore, the action space for a classification task assigning labels $y \in \mathcal{Y}$ to inputs $\mathbf{x} \in \mathcal{X}$ is $\mathcal{A} = \{f_\theta : \mathcal{X} \to \mathcal{Y}\}$. The risk function is in turn a function $\mathcal{R} : \mathcal{F} \times \mathcal{A} \to \mathbb{R}^+$, which is the loss incurred by taking an action in $\mathcal{A}$ (making a prediction with a specific configuration of neural parameters) when the 'state of nature', the true function, is $f \in \mathcal{F}$. In the case of MAML, this is represented by the language-specific inner loop loss $\mathcal{L}_{\mathcal{T}_i}(\cdot)$ in equation (7.9).

The decision for the optimal action given the sample space, the function $\delta : \mathcal{X} \times \mathcal{Y} \to \mathcal{A}$, is usually determined via gradient descent optimisation for a neural network. The optimal action, however, may vary depending on the language, which results in multiple possible 'states of nature'. Usually, there is no procedure δ whose loss is inferior to all others, such that:

$$\nexists \delta \; \mathcal{L}(\mathcal{T}_i, \delta) < \mathcal{L}(\mathcal{T}_i, \delta') \; \forall \mathcal{T}_i \in \mathcal{T}, \delta \neq \delta' \tag{7.10}$$

Therefore, decision functions have to be compared based on a global criterion rather than in a pair-wise fashion between languages. As previously anticipated, equation (7.9) minimizes the *expected* risk across languages, for an arbitrary choice of prior $p(\mathcal{T})$. In decision theory, a decision $\delta^\star$ with this property is called *Bayes criterion*.

7.4.2 *Alternative Criteria*

There exist alternative criteria to the Bayes criterion that are more justified in a setting that entails transfer between non-i.i.d. domains. Rather than minimising the Bayes risk, in this work, we propose to adjust MAML to either minimise the maximum risk (minimax criterion) or to enforce constraints on the risk for a subset of languages (Neyman–Pearson criterion). This is likely to yield more robust predictions for languages that are outliers to the training distribution. As demonstrated in section 7.3, this definition encompasses most of the world's languages.

7.4.2.1 *Minimax Criterion*

Rather than the expected risk, the criterion could depend instead on the worst case scenario, i.e. the language for which the risk is *maximum*. This requires to select such a language with max. As an alternative to reinforcement learning (Zhang et al., 2020a), to keep our model fully differentiable, we relax the operator by treating the choice of language as a categorical distribution $\mathcal{T}_i \sim \text{Cat}(\cdot \mid \boldsymbol{\tau})$. The parameters $\boldsymbol{\tau} \in [0,1]^{|\mathcal{T}|}, \sum_i \tau_i = 1$ consist of language probabilities and are learned in an adversarial fashion:

$$\min_{\theta} \max_{\mathcal{T}_i \sim \text{Cat}(\cdot|\boldsymbol{\tau})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \eta \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta, \mathcal{D}_{\text{train}})}, \mathcal{D}_{\text{val}}) \tag{7.11}$$

Equation (7.11) can be interpreted as a two-player game between us (the scientists) and nature. We pick an action θ. Then nature picks a language $\mathcal{T}_i \in p(\mathcal{T})$ for which the risk is maximum given our chosen action. Therefore, our goal becomes to minimise such risk.

7.4.2.2 *Neyman–Pearson Criterion*

As an alternative, we might consider minimising the expected risk, but subject to a guarantee that the risk does not exceed a certain threshold for a subset of languages. In practice, we may want to enforce a set of inequality constrains, so that we minimise equation (7.9) subject to $\{\mathcal{L}_{\mathcal{T}_i} \leqslant r \;\forall \mathcal{T}_i \in \mathcal{C}\}$, where $r \in \mathbb{R}_+$ is a hyper-parameter. In general, $\mathcal{C} \subseteq \mathcal{T}$ can be any subset of the training languages; in practice, here we take $\mathcal{C} = \mathcal{T}$. Constrained optimisation is usually implemented through Lagrange multipliers, where we add as many new terms to the objective as we have constraints (Bishop, 2006, ch. 7):

$$\begin{aligned} &\min_{\theta} \max_{\lambda} \sum_{\mathcal{T}_i} \frac{1}{|\mathcal{T}|} \mathcal{L}_{\mathcal{T}_i} + \sum_{\mathcal{T}_i} \lambda_i (\mathcal{L}_{\mathcal{T}_i} - r) \\ &= \min_{\theta} \max_{\lambda} \sum_{\mathcal{T}_i} \left(\frac{1}{|\mathcal{T}|} + \lambda_i \right) \mathcal{L}_{\mathcal{T}_i} - \lambda_i r \end{aligned} \tag{7.12}$$

where $\boldsymbol{\lambda}$ is a vector of non-negative Lagrange multipliers $\{\lambda_i \geqslant 0 \quad \forall \lambda_i \in \boldsymbol{\lambda}\}$ to be learned together with the parameters θ, but adversarially.

Intuitively, if the risk for the estimated parameters θ lies in the permissible range, the constraints should become inactive $\{\lambda_i = 0 \quad \forall \lambda_i \in$

$\boldsymbol{\lambda}\}$, i.e. each Lagrange multiplier should go towards 0. Otherwise, the solution should be affected by the constraints, which should keep θ from trespassing the boundary $\{\mathcal{L}(\theta)_{\mathcal{T}_i} = r \quad \forall \mathcal{T}_i \in \mathcal{T}\}$. In gradient-based optimisation, this unfolds as follows: the gradient of each λ_i depends uniquely on $(\mathcal{L}_{\mathcal{T}_i} - r)$. Due to being maximised, the value of each λ_i increases when the corresponding risk is above the threshold, and shrinks otherwise. Incidentally, note that the Lagrangian multipliers at the critical point $\theta^\star$ are equal to the negative rate of change of r, as $\frac{\partial \mathcal{R}(\theta^\star)}{\partial r} = -\boldsymbol{\lambda}$. In other words, upon convergence λ_i expresses how much we can decrease the risk in $\mathcal{T}_i$ as we increase the threshold.

CONSTRAINED PARAMETERS The additional variables $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$, contrary to the neural parameters, are constrained in the values they can take. In neural networks, there are two widespread approaches to coerce variables within a certain range, viz. reparametrisation and gradient projection (Beck and Teboulle, 2003).[2] For simplicity's sake, we opt for the former, which just requires us to learn unconstrained variables and scale them with the appropriate functions. Thus, we redefine the above-mentioned variables as $\boldsymbol{\tau} \triangleq \text{softmax}(\boldsymbol{\tau}_u)$ and $\boldsymbol{\lambda} \triangleq \text{softplus}(\boldsymbol{\lambda}_u)$.

7.5 OPTIMISATION IN 2-PLAYER GAMES

Based on the formulation of Minimax MAML and Neyman–Pearson MAML in section 7.4.2, both are evidently instances of two-player games. On one hand, the first agent minimises the risk with respect to θ; on the other, the second agent maximises the risk with respect to $\boldsymbol{\tau}$ (for minimax) or $\boldsymbol{\lambda}$ (for Neyman–Pearson). In other words, both optimise the same (empirical risk) function in equation (7.11) or equation (7.12), respectively, but with opposite signs. However, the first term of equation (7.12) does not depend on $\boldsymbol{\lambda}$. Therefore, Minimax MAML is a zero-sum game, but not Neyman–Pearson MAML.

If the risk function were convex, the solution would be well-defined as the Nash equilibrium. But this is not the case for a non-linear function such as a deep neural network. Therefore, we resort to an approximate solution through optimisation. The simplest approach in this scenario is Gradient Descent Ascent (GDA), where the set of parameters of both players are optimised simultaneously through gradient descent for the first player and gradient ascent for the second player. With a slight abuse of notation, let us define $\mathcal{R} \triangleq \mathcal{R}(\theta_t, \boldsymbol{\alpha}_t)$, where $\boldsymbol{\alpha}_t$ stands for the adversarial parameters ($\boldsymbol{\tau}_t$ for Minimax and $\boldsymbol{\lambda}_t$ for Neyman–Pearson) at time t. Then the update rule is:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{R} \tag{7.13}$$

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \eta \nabla_{\boldsymbol{\alpha}} \mathcal{R} \tag{7.14}$$

for a learning rate $\eta \in \mathbb{R}$. Equations (7.13) and (7.14) are equivalent to allowing each player to ignore the other's move and act as if it will remain stationary. This naïve assumption often leads to divergence

2 https://vene.ro/blog/mirror-descent.html

or sub-par solutions during optimisation (Schäfer and Anandkumar, 2019).

7.5.1 *Symplectic Gradient Adjustment*

To overcome the limitations of GDA, several independent works (Balduzzi et al., 2018; Gemp and Mahadevan, 2018; Letcher et al., 2019) proposed to correct equations (7.13) and (7.14) with an additional term. This consists of a matrix-vector product between the mixed second-order derivatives ($D^2_{\theta\alpha}\mathcal{R}$ and $D^2_{\alpha\theta}\mathcal{R}$, respectively)[3] and the gradient of the risk with respect to the adversarial parameters ($\nabla_\alpha\mathcal{R}$ and $\nabla_\theta\mathcal{R}$, respectively). The resulting optimisation algorithm, Symplectic Gradient Adjustment (SGA), updates parameters as follows:

$$\theta_{t+1} = \theta_t - \eta\nabla_\theta\mathcal{R} - \eta^2 D^2_{\theta\alpha}\mathcal{R}\,\nabla_\alpha\mathcal{R} \tag{7.15}$$

$$\alpha_{t+1} = \alpha_t + \eta\nabla_\alpha\mathcal{R} - \eta^2 D^2_{\alpha\theta}\mathcal{R}\,\nabla_\theta\mathcal{R} \tag{7.16}$$

Intuitively, the mixed second-order derivative represents the interaction between the players, and the adversarial gradient represents the opponent's move if they follow the simple GDA strategy. Schäfer and Anandkumar (2019) cogently demonstrate how equations (7.15) and (7.16) correspond to an approximation of the Nash equilibrium[4] of a local bi-linear approximation (with quadratic regulariser) of the underlying game dynamics.

In practice, estimating the above-mentioned products is tedious because of their space and time complexity. Therefore, we resort to an approximation known as *Hessian-vector product* (Pearlmutter, 1994). For the third term of equation (7.15):

$$\begin{aligned} &D^2_{\theta\alpha}\mathcal{R}(\theta,\alpha)\nabla_\alpha\mathcal{R}(\theta,\alpha) \\ =&\frac{\partial}{\partial h}\nabla_\theta\mathcal{R}(\theta,\alpha + h\nabla_\alpha\mathcal{R}(\theta,\alpha))\Big|_{h=0} \end{aligned} \tag{7.17}$$

And similarly for the matrix product term in equation (7.16), by swapping θ and α in equation (7.17).

7.5.2 *Adaptive Learning Rate and Momentum*

While SGA may provide a more appropriate optimisation framework for competitive games, it still lacks several defining features of optimisers that accelerate convergence, such as first-order momentum and adaptive learning rate (second-order momentum). Therefore, we modify the update rule in equations (7.15) and (7.16) to include both of these. Our starting point is Adam (Kingma and Ba, 2015). The changes we apply are the following (also illustrated in algorithm 3):

3 Here $D^2_{wz}\mathcal{R}$ stands for the sub-matrix of the Hessian containing the derivative of the risk taken first with respect to w and then with respect to z.

4 A Nash equilibrium is a pair of strategies whose unilateral modification cannot result in loss reductions.

Algorithm 3 Adaptive Symplectic Gradient Adjustment (ASGA)

Require: $\eta \in \mathbb{R}_+$: Learning rate
Require: $\beta_1, \beta_2 \in [0, 1)$: Decay rates
Require: θ_0, α_0: Initial parameter values
Require: $\mathcal{R} \triangleq \mathcal{R}(\theta_{t-1}, \alpha_{t-1}) : \mathbb{R}^{|\theta|+|\alpha|} \to \mathbb{R}$
1: $m_0 \leftarrow 0$ Initialise first moments
2: $v_0 \leftarrow 0$ Initialise second moments
3: $t \leftarrow 0$ Initialise time step
4: **while** θ_t, α_t not converged
5: $\quad t \leftarrow t + 1$
6: $\quad g_{\theta,t} \leftarrow \nabla_\theta \mathcal{R} + \eta D_{\theta\alpha} \mathcal{R} \, \nabla_\alpha \mathcal{R}$
7: $\quad g_{\alpha,t} \leftarrow \nabla_\alpha \mathcal{R} - \eta D_{\alpha\theta} \mathcal{R} \, \nabla_\theta \mathcal{R}$
8: $\quad g_t \leftarrow g_{\theta,t} \oplus g_{\alpha,t}$
9: $\quad m_t \leftarrow \beta_1 \, m_{t-1} + (1 - \beta_1) \, g_t$
10: $\quad v_t \leftarrow \beta_2 \, v_{t-1} + (1 - \beta_2) \, g_t^2$
11: $\quad \hat{m}_t \leftarrow m_t \, / \, (1 - \beta_1^t)$
12: $\quad \hat{v}_t \leftarrow v_t \, / \, (1 - \beta_2^t)$
13: $\quad \theta_t \leftarrow \theta_{t-1} - \eta \cdot \hat{m}_{\theta,t} \, / \, (\sqrt{\hat{v}_{\theta,t}} + \epsilon)$
14: $\quad \alpha_t \leftarrow \alpha_{t-1} + \eta \cdot \hat{m}_{\alpha,t} \, / \, (\sqrt{\hat{v}_{\alpha,t}} + \epsilon)$
15: **end while**
16: **return** θ_t, α_t

1. The current difference (lines 6–7) is adjusted with the terms introduced in equations (7.15) and (7.16) by Schäfer and Anandkumar (2019).
2. The exponentially decayed, unbiased estimates of the expectations over mean and standard deviation are computed similarly to Adam. However, note that, in line 14, the update of the adversarial parameters corresponds to an ascent (rather than a descent).

This results in a novel optimiser, Adaptive Symplectic Gradient Adjustment (ASGA). We employ ASGA in our experiments to optimise the objectives of Minimax MAML and Neyman–Pearson MAML, as it enables a fair comparison with Adam-optimised Bayes MAML.

7.6 EXPERIMENTS

We now outline the main experiments of our work on multilingual NLP. We evaluate our methods on part-of-speech (POS) tagging, a sequence labelling task, and question answering (QA), a natural language understanding task.

We focus on POS given its ample coverage of languages and its frequent use as a benchmark for resource-poor NLP (Das and Petrov, 2011; Ponti et al., 2021). In fact, cross-lingual transfer in sequence labelling tasks was demonstrated to be the most challenging, as knowledge of linguistic structure is more language-dependent than semantics (Hu et al., 2020). However, we also include QA to illustrate the generality of our methods for cross-lingual NLP. In this task, given

the gold passage and a question, the system has to predict the beginning and end positions of a single contiguous span containing the answer.

Data. POS data are sourced from the Universal Dependencies (UD) treebanks[5] (Zeman et al., 2020) and QA data from the 'gold passage' variant of TyDiQA (Clark et al., 2020).[6] We retain the original training, development, and evaluation sets of UD. In TyDiQA, we use the original development set for evaluation.[7] For meta-learning, $\mathcal{D}_{\texttt{train}}$ and $\mathcal{D}_{\texttt{val}}$ examples are both obtained from disjoint parts of the training set.

We aim to create a partition of languages between training and evaluation that corresponds to the most realistic scenario in deploying NLP technology on resource-poor languages spoken around the world. Therefore, we reserve for evaluation all language isolates and languages with at most 2 family members in each dataset. We use all the remaining languages in the dataset for training. Therefore, for POS, the evaluation set spans 16 treebanks (14 languages, 11 families) and the training set 99 treebanks; QA comprises 9 languages (7 families). We hold out 4 of them in turn for evaluation (except English) and use the rest for training. We provide the full list of languages in appendix A.4.1.

Training. In all tasks, we train a neural network consisting of two stacked modules: an encoder and a classifier. The encoder is a 12-layer, 768-hidden unit, 12-head Transformer initialised with multilingual BERT Base (mBERT), which was pre-trained on cased text from 104 languages.[8] The classifier is a single affine layer for TyDiQA and a 2-layer Perceptron (with 1024 hidden units) for POS tagging. The combined parameters of the encoder and classifier correspond to θ from section 7.4.

These are meta-learned via Meta-SGD (Li et al., 2017), a first-order MAML variant where each parameter is assigned a separate inner-loop learning rate η. Moreover, each η is trained end-to-end based on the outer-loop loss (such as equation (7.9) for the Bayes criterion).[9] Similar to Bansal et al. (2020), to avoid an explosion in the number of parameters, we assign a per-layer learning rate (rather than per-parameter). To avoid overfitting, we employ both dropout (with a probability of 0.2) and early stopping (with a patience of 10). For the Neyman-Pearson formulation, we set $r = 0.1$ as a threshold for all language-specific losses.[10] The parameters $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ were initialized uniformly as $\frac{1}{|\mathcal{T}|}$. Complete details of the hyper-parameters for all settings are given in Appendix A.4.2.

5 https://universaldependencies.org/
6 https://github.com/google-research-datasets/tydiqa
7 This is necessary as we need to access this set to simulate few-shot learning, but the original evaluation set is not public.
8 https://github.com/google-research/bert/blob/master/multilingual.md
9 We implement Meta-SGD with the `learn2learn` package (Arnold et al., 2020).
10 We also experimented with a dynamic threshold which corresponded to the average language-specific loss of the last 10 episodes. However, this yielded sub-par results.

k ▷	0	5	10	20
		F_1 Score		
J	51.01	62.96±2.5	66.00±1.9	68.66±1.7
B	51.50	63.87±2.8	67.03±2.1	69.46±1.8
MM	51.82	63.67±2.7	66.88±2.0	69.55±1.8
NP	51.68	63.84±2.9	67.13±2.1	69.65±1.9
MM+	52.46	**64.71±2.9**	**67.89±2.3**	**70.25±2.0**
NP+	**53.05**	64.26±2.6	67.57±2.1	69.98±1.9

Table 14: F_1 scores for POS tagging in UD across different *k*-shots. We report the mean and standard deviation across 16 treebanks.

Methods. To assess the effectiveness of the proposed criteria and optimisers, we compare them with two competitive baselines, while maintaining the same underlying neural architecture: (i) **J**: a joint multi-source transfer method where a model is trained on the concatenation of the datasets for all languages; (ii) **B:** the original MAML (Finn et al., 2017) with Bayes criterion and uniform prior. Our choice of baselines is justified by the fact that these methods (or variations thereof) are currently state of the art for the tasks of POS and QA, as well as other innumerable NLP applications (Nooralahzadeh et al., 2020; Ponti et al., 2021; Wu and Dredze, 2019). In addition, we evaluate the following combinations: (iii) **MM:** MAML with a minimax criterion, optimised with GDA; (iv) **NP:** MAML with a Neyman–Pearson (constrained) criterion, optimised with GDA; (v) **MM+:** MAML with a minimax criterion, optimised with ASGA; and (vi) **NP+:** MAML with a Neyman–Pearson criterion, optimised with ASGA.

Evaluation. For each evaluation language in a given task, we randomly sample $k \in \{0, 5, 10, 20\}$ examples from the evaluation data as the support set (for adaptation) and the rest of the examples as the query set (for testing). When $k > 0$, we repeat the evaluation 100 times and report the following average metrics: (i) F_1 score for POS tagging, and (ii) exact-match (EM) and F_1 scores for QA.[11] Due to lack of space, we only report the average mean and standard deviation across languages for each model described above.

7.7 RESULTS AND DISCUSSION

We report the results for POS tagging in table 14 and for QA in table 15. These include mean and standard deviation across languages. Note that, in this case, the standard deviation is by no means an interval for statistical significance, but rather reflects the heterogeneity among the evaluation languages. In what follows, we address a series of questions in the light of these figures.

Baselines. MAML and joint multi-source transfer are both strong contenders as state-of-the-art methods for cross-lingual transfer, but

11 We refer the reader to Rajpurkar et al. (2016a) for a precise definition of these metrics.

k ▷	0	5	10	20
Exact Match				
J	46.76	49.53±3.7	51.54±2.9	**53.51±2.4**
B	46.60	48.41±3.4	50.24±2.9	52.02±2.6
MM	**48.33**	**50.08±3.4**	**51.68±2.9**	**53.49±2.4**
NP	46.71	49.24±3.3	50.95±2.9	52.76±2.4
MM+	46.87	47.74±3.8	49.42±3.4	51.40±2.5
NP+	48.02	48.77±3.9	50.75±3.1	52.66±2.6
F_1 Score				
J	61.66	63.75±3.3	65.39±2.3	67.01±1.9
B	62.51	63.29±3.2	64.87±2.5	66.31±2.1
MM	**63.06**	**64.37±3.1**	**65.83±2.6**	**67.45±2.1**
NP	61.89	63.84±2.9	65.23±2.6	66.88±1.9
MM+	62.10	62.63±3.2	64.11±2.9	65.89±2.1
NP+	62.75	62.98±3.6	64.77±2.9	66.57±2.2

Table 15: Results for QA in TyDiQA across different *k*-shots. We report the mean and standard deviation across 8 languages of the exact match score (above) and the F_1 score (below).

which one is better? By comparing J and B rows, no definite response emerges in our experiments. While MAML outperforms its competitor in POS tagging, it lags behind in QA. We speculate that the larger pool of training languages available in POS tagging (22 times more than QA) endows meta-learning with better generalisation capabilities. Both methods, however, surpass single-source transfer from English SQuAD (Rajpurkar et al., 2016a) in the zero-shot setting by a large margin: Clark et al. (2020) report 56.4 F_1 score in average for TyDiQA, which is 6.66 points below our best model.

Criteria. The minimax and Neyman-Pearson criteria both improve over the Bayes criterion baseline, although the latter more sporadically. Compared to the B rows, MM+ achieves gains for every k in POS tagging, with 0.94 points of margin at $k = 0$ and 0.79 at $k = 20$. The same holds for MM in QA, with margins that span from 1.73 at $k = 0$ to 1.47 at $k = 20$ in the Exact Match metric, and from 0.55 at $k = 0$ to 1.14 at $k = 20$ in F_1 score. Therefore, Minimax MAML is remarkably consistent in outperforming the baselines, although the gains are sometimes significant, sometimes only marginal. This is also reflected in language-specific performances, available in tables 26 and 27 and tables 28 and 29 in the Appendix. For POS tagging, the F_1 scores of only 2 languages (Indonesian and Naija) moderately decrease, whereas the rest of the 14 languages show improvements.

Incidentally, it may be worth noting that we did not perform any large-scale search over hyper-parameters like $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ initialisations, the threshold r, or differential learning rates for maximised and minimised parameters. Therefore, these early results are amenable to improve even further in the future. This lends credence to our proposi-

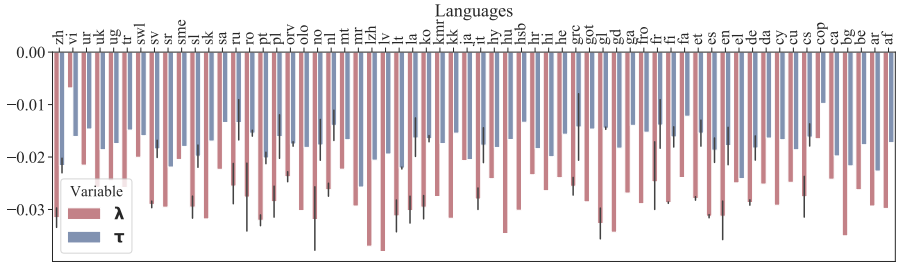


Figure 24: Unconstrained values of τ_u and λ_u upon convergence in MM+ and NP+ models for POS tagging.

k ▷	0	5	10	20
		F_1 Score		
J	14.34	33.32	37.52	40.83
B	24.11	35.03	40.38	44.92
MM	20.41	37.61	43.00	45.83
NP	**26.81**	39.23	42.70	45.25
MM+	16.42	37.41	43.57	45.21
NP+	22.55	**39.95**	**45.41**	**48.12**

Table 16: The minimum F_1 scores of our models across languages, for POS tagging.

tion that minimax and Neyman–Pearson criteria are more suited for out-of-distribution transfer to outlier languages.

Optimiser. The results for the proposed optimiser ASGA (algorithm 3) are favourable in comparison to Gradient Descent Ascent via Adam (Kingma and Ba, 2015) for POS tagging; on the other hand, the opposite trend is observed for QA. Therefore, future investigations are required to shed further light on modifications such as the Symplectic Gradient Adjustment. A tentative explanation of such discrepancy could be the disproportionate number of training languages available in either task.

To get insights into the game dynamics of the adversarial criteria, we plot the unconstrained values for τ_u and λ_u upon convergence in figure 24. Interestingly, both variables appear to follow the same profile of peaks and troughs; therefore, as expected, languages chosen adversarially in MM have also higher Laplace multipliers in NP. To this group belong for instance languages with rare scripts (e.g. Coptic) or with no relatives in the training languages (e.g. Vietnamese). As a final note, we remark that the proposed criteria and optimiser are in principle more general than NLP and could facilitate transfer in other fields. While this thread of research transcends the scope of our work, we illustrate an example for regression in appendix A.4.3.

Minimum Scores across Languages. In addition to the *average* cross-lingual performance, we also report the *minimum* cross-lingual performance for POS tagging in table 16 and for QA in table 17. This corresponds to the lowest score achieved across all evaluation languages.

k ▷	0	5	10	20
Exact Match				
J	42.33	**45.97**	47.22	**49.47**
B	**42.75**	44.58	46.44	48.24
MM	41.01	45.33	**47.59**	49.21
NP	40.39	44.89	46.40	48.80
MM+	41.01	41.87	45.32	47.11
NP+	37.44	42.92	46.30	48.88
F_1 Score				
J	52.43	59.27	59.88	62.11
B	51.10	**59.60**	60.64	62.10
MM	53.10	59.21	**61.86**	63.43
NP	52.83	59.31	60.03	61.84
MM+	51.93	57.91	59.86	61.52
NP+	**53.96**	57.21	61.74	**63.55**

Table 17: The minimum Exact Match and F_1 scores of our models across languages, for QA.

For POS tagging, we observe that NP and NP+ outperform J and B by 7-12 and 2-5 F1 points, respectively. This reveals that worst-case and constrained risk minimisation drastically uplifts the scores for the most disadvantaged language. Nevertheless, the opposite trend is observed for QA: MM(+) and NP(+) do not alter the minimum score with respect to the F_1 metric, and even degrade it with respect to the exact-match metric. Again, we conjecture that these mixed findings may depend on the different amount and distribution of the training languages in the corresponding datasets: UD offers greater language coverage than TyDiQA, which gives better guidance.

7.8 RELATED WORK

MAML is a cutting-edge method for cross-lingual transfer in several NLP tasks (Gu et al., 2018; Li et al., 2020b; Nooralahzadeh et al., 2020; Wu et al., 2020a, *inter alia*). However, in all these experiments, the model is adopted in its standard formulation, minimising the expected risk. Therefore, its performance is prone to suffer in outlier languages. Moreover, the assumptions underlying our proposed variants are different from other instances of robust optimisation in NLP Globerson and Roweis, 2006; Oren et al., 2019. In particular, the target language distributions are not explicitly treated as subspaces or covariate shifts of source languages. In separate fields such as vision, previous attempts at worst-case-aware meta-learning include Collins et al. (2020), who use a Euclidean version of the robust stochastic mirror-prox algorithm, and Wang et al. (2020c), who rely on reinforcement learning. Our formulation is both fully differentiable and broader, as the decision-theoretic interpretation admits alternative cri-

teria for MAML. What is more, to our knowledge we are the first to successfully augment MAML with minimax criteria in cross-lingual NLP and with Neyman–Pearson criteria in general.

7.9 CONCLUSIONS

To perform cross-lingual transfer to low-resource languages, under a decision-theoretic interpretation Model-Agnostic Meta-Learning (MAML) minimises the expected risk across training languages. Generalisation then relies on the evaluation languages being identically distributed. However, this assumption is incongruous for cross-lingual transfer in realistic scenarios. Therefore, we propose more appropriate training objectives that are robust to out-of-distribution transfer: Minimax MAML, where worst-case risk is minimised by learning an adversarial distribution over languages; and Neyman–Pearson MAML, where constraints are imposed on language-specific losses, so that they remain below a certain threshold. From a game-theoretic perspective, both of these variants consist of 2-player competitive games. Therefore, we also explore adaptive optimisers that take into account the underlying game dynamics. The experimental results on zero-shot and few-shot learning for part-of-speech tagging and question answering, whose datasets span tens of typologically diverse languages, confirm that in several settings the proposed criteria are superior to both vanilla MAML and transfer from multiple source languages.

8 CONCLUSIONS

This thesis looked at different aspects of neural coreference resolution to improve the performance of already trained models without using labeled data and without increasing model capacity. Three methods are proposed, each of which approaches the problem from a different direction. While one method (Chapter 3) uses data annotated for other tasks, the other two (Chapters 2, 4) only use free-form unannotated text.[1] All of them lead to improvements in performance. We conclude this work by summarizing the key takeaways from each chapter.

Chapter 2 introduced a reinforcement learning-based finetuning technique that used an external knowledge base to verify whether the resolution made by a model is accurate or not. It also provided a way to generalize the knowledge present in knowledge graphs by using scorers inspired by Universal Schema. Though we applied this method only to coreference resolution, it is easily extendable to other tasks which require outputs to be faithful to a knowledge source.

In Chapter 3, we transformed anaphora resolution tasks to a form resembling cloze-style question-answering. This allowed us to adapt existing QA models and datasets for training our tasks. Experiments show that such *task recasting* is helpful, especially in cases where tasks do not have large training sets, like sluice and verb-phrase ellipsis resolution. With this method, both ellipsis tasks saw huge gains resulting in a new state of the art.

Chapter 4 introduced a joint learning method that finetuned multiple models using coherence rewards. We combined semantic role labels and coreference links to form simple semantic graphs whose coherence provides a supervision signal that can be exploited using reinforcement learning. We also showed that this method is robust to model size and works well, especially for smaller models. It should be noted that this method neither uses auxiliary annotated data nor an external knowledge base.

In Chapter 5, we look at coreference annotation through the lens of entity linking. Instead of linking spans of text with each other, we propose an annotation scheme that enables us to link pro-forms with entities in a knowledge graph. According to our analysis, this method not only decreased the annotation times it also increased inter-annotator agreement scores. In general, this leads to better quality datasets.

In Chapter 6, we introduced *Focus Attention*, a mechanism that biases the decoder of a transformer-based seq2seq model to generate thematically relevant text. We showed that the inclusion of this type of attention results in abstractive summarization systems becoming more faithful to their inputs. This attention formulation also al-

1 Note that the method described in Chapter 2 also uses an external knowledge base.

lowed us to create a new *Focus Sampling* technique which enables the model to generate diverse outputs while maintaining its faithfulness. Though we evaluate our method only on summarization, we argue it is broadly applicable to many generation tasks.

Finally, in Chapter 7, we introduced two new variants of MAML, which break the i.i.d assumption ingrained in the original. Since these methods can be seen as 2-player games, we also introduce a new trick to efficiently optimize them with a *sympletic gradient adjustment*. We showed that the alternative criteria help out-of-distribution languages, especially in zero- and few-shot POS-tagging and QA. Our analysis suggests that the risk threshold is an important hyperparameter for Neyman-Pearson MAML to work well and that in general, Minimax MAML is more effective.

FUTURE DIRECTIONS With the introduction of transformer architectures with linear or quasi-linear complexities, we can now encode long-form documents without having to split them into chunks (Wang et al., 2020b; Zaheer et al., 2020). Since they are pre-trained on large amounts of data, they become very good at identifying and resolving anaphora. This has resulted in lethargy in the coreference resolution space. However, even these models cannot precisely detect cross-document (CD) coreference. Today, entity-based and event-based CD coreference resolution has become especially important for tasks like fact-checking, multi-document summarization, multi-hop QA, etc. This line of work has only recently gained attention, and we already see works like Cattan et al., 2020, which propose to standardize the evaluation and other aspects of the task. ECB+ (Cybulska and Vossen, 2014) has emerged as a standard dataset for CD coreference, with works also evaluating on multi-hop QA datasets like the one proposed by Khashabi et al., 2018.

An interesting tangential line of work is in the domain of evaluation metrics. As seen in Chapters 2 and 5 of this thesis, the current metrics do not reflect the performance of the models in cases where the models detect and link mentions which are not annotated in the evaluation set. Since manual annotation is hard and error-prone, augmenting them with automatic metrics may provide a better insight into model performance. Finally, it may also be worthwhile to explore the possibility of biasing self-attention layers to identify entity chains more accurately.[2] This would not only improve discourse representations, but it would also eliminate the need for pipeline-based approaches where coreference resolution is performed as a preprocessing step before the actual task, and allow us to train models end-to-end.

2 This idea is not new. It has already been incorporated into recurrent networks in Dhingra et al., 2018.

Part IV

APPENDIX

A

APPENDIX

A.1 CHAPTER 3

A.1.1 *Similarity between Ellipsis and Coreference Resolution*

Linguists have long pointed out deep links among different forms of ellipsis, as well as between ellipsis and pronominal anaphora. For example, Merchant (2001) presents a unified account of ellipsis phenomena within a minimalist syntactic framework, and theorists such as Postal (1966) and Elbourne (2013) go so far as to argue that pronouns are also elliptical forms. The exact nature of the connections between ellipsis and anaphoric constructions remains a subject of controversy among linguists. However it is clear that there are rooted connections, and in our view these connections represent potential areas to be exploited with forms of knowledge transfer among datasets of different types.

Typically in NLP, ellipsis and coreference have been treated as distinct tasks. Possible exceptions include Lin et al. (2016), who present a rule-based, feature-rich system for handling ellipsis and coreference in Chinese medical dialogues, but the synergy between the two subsystems is limited; and Banjade et al. (2015), who reduce ellipsis and coreference to problems of alignment to an auxiliary text implicitly describing the universe of the dialogue in question.

A.1.2 *QA Models*

We briefly describe the architectures of the QA models below. All experiments are conducted on a single 12 GB GPU. For all models, we use the hyperparameter values recommended in their respective papers.

DRQA The Document Reader component of DrQA consists of a context and a question encoder followed by two span prediction classifiers. The context encoder is a multi-layer bi-directional LSTM (Hochreiter and Schmidhuber, 1997a) which takes in word embeddings (Pennington et al., 2014a, GloVe), similarity based features (whether the token appears in the question in it's original, lowercase or lemma form), and other token level features (positional tags, named entities and term frequency) as input. The concatenation of each layer's hidden units is used as the context vector. The question encoder is another LSTM which takes word embeddings as input and combines the resulting hidden units using a simple attention mechanism to form the question vector. A bilinear term which captures the similarities between context and question vectors is used to combine the

two vectors and the resulting vector used as input to the span prediction classifiers. The two classifiers predict the start and the end span respectively and are trained independently.

QANET In QANet, each encoder layer is a stack of depthwise separable convolutions followed by a multi-head self-attention mechanism placed inside a residual block. Initially, words in the context and question are embedded using a combination of GloVe and character embeddings. They are then contextualized with an encoder block. The representations are then passed through a context-query attention layer to obtain a combined representation of the context and question. This is further passed through three encoding blocks and the final output is input to a classifier for predicting the answer spans.

BERT We use the pre-trained $\text{BERT}_{\text{BASE}}$ uncased model to encode questions and their contexts. It has 12 Transformer blocks, 12 self-attention heads, and a hidden size of 768. Word piece tokenization (Wu et al., 2016) is performed, both on the context paragraph and the question. The boundaries of the two sequences are marked by dummy symbols. The context and the question are joined with a [SEP] token in between, and the [CLS] token is prepended at the beginning to form the input. The representation of the [CLS] token is fed into a single-layer MLP with 2 outputs which is used to predict the span indices.[1]

A.1.3 *Coreference Resolution*

In this section, we analyse the best performing coreference models and discuss why they cannot be compared with other works in literature.

A.1.3.1 *Error Analysis*

The JOINT OntoNotes model improves a little over the SINGLE-TASK counterpart. Here we examine specific referential forms in OntoNotes (WikiCoref has similar traits), as shown in Figure 25. In general, performance is better on frequent pronouns – e.g., 'he' over 'she', 'this' and 'that'. An exception to this is that 'it' is less accurate, but more frequent than 'he'. It is notable that the possessive pronouns ('his', 'her', 'its') are all more accurate than their nominative counterparts ('he', 'she', 'it'), perhaps because they tend to have a closer connection to their antecedents. Overall, the single-word referential forms are less accurate than multiple-word forms. For example, definite descriptions (forms beginning with 'the') are more accurate than any of the single-word forms, with the exception of 'its'. We speculate that multi-word forms provide more specific information, thus limiting the set of potential antecedents. It is also interesting to break down error by the grammatical gender of the pronouns. Male pro-

1 We use the implementation detailed in Wolf et al. (2019).

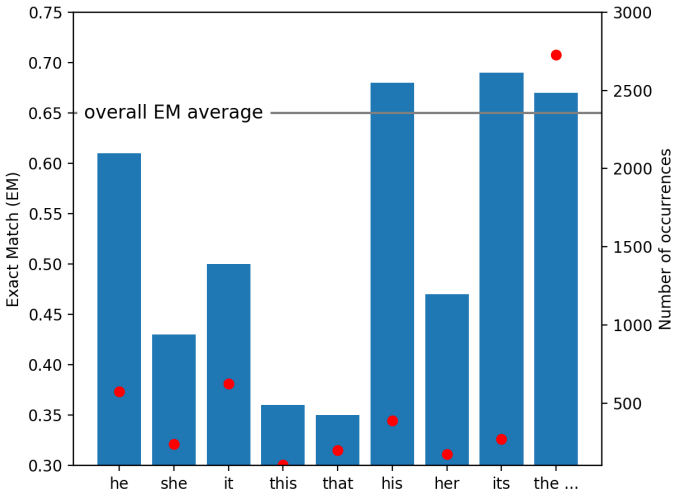


Figure 25: Exact match percentage (bars) and number of occurrences (dots) of referential forms in OntoNotes

nouns generally tend to be more accurate than their female counterparts. Antecedents of 'he' and 'his' are matched 20% more frequently than for 'she' and 'her'. This is probably due to an unfortunate bias in OntoNotes, where female pronouns are 50% rarer than male pronouns.

A.1.3.2 *Result Comparability*

Converting coreference into QA fundamentally changes the coreference resolution problem: It, on the one hand, makes the coreference resolution problem harder, in that we require the identification of a specific antecedent span, rather than any mention in the entity chain; on the other hand, the problem becomes easier by providing the bracketing of the mention that needs to be resolved. Due to these differences, it is not possible to directly compare our results with others in literature. For analysis, to make our results more comparable with Lee et al. (2018c), we provided their model with the bracketing of the mentions and considered the first mention to be the antecedent. This way we can reinterpret their clusters as question-answer pairs and not penalize them for mention bracketing errors, only considering pairs where they correctly identify mentions. Note this gives their model an advantage over ours, as their model considers multiple sources of evidence for inferring the coreference links, and gets to pick the subset of data on which the models are compared. On OntoNotes, in this setting, and after pruning around 7,358 mentions Lee et al. (2018c) bracketed wrongly, their new average F_1 score is 75.9. Our performance on the same subset of the data is 72.1. Upon manual inspection, we see the model in Lee et al. (2018c) has a strong bias favoring nominal antecedents, whereas our model is more likely to predict clausal antecedents. On WikiCoref, our model remains better than the previous state of the art by some margin, with an F_1 of 69.2 over 43.6.

	ON		PC		PD-G		PD-W		WC		WB		Avg.	
	B.	O.	B.	O.	B.	O.	B.	O.	B.	O.	B.	O.	B.	O.
1	62.6	62.6	42.4	44.4	47.0	46.7	32.7	33.2	39.8	39.8	69.6	69.7	49.0	**49.4**
2	61.0	61.4	42.3	45.4	48.4	48.9	37.7	37.7	45.2	45.0	63.7	67.2	49.7	**50.6**
3	64.2	64.2	44.8	46.1	50.0	50.3	40.2	40.3	49.7	49.4	66.8	67.0	52.6	**52.8**
4	65.8	66.4	44.3	47.5	51.3	52.1	42.4	42.1	50.3	50.6	62.5	64.7	52.8	**53.9**
5	68.5	68.7	45.9	48.5	54.7	54.7	42.4	41.8	54.1	54.2	68.6	69.2	55.7	**56.2**
6	71.5	71.5	46.9	48.6	56.9	57.3	44.8	44.2	55.0	55.3	71.6	72.2	57.8	**58.2**

Table 18: COREFERENCE RESOLUTION results of single-task models.

	ON		PC		PD-G		PD-W		WC		WB		Avg.	
	B.	O.	B.	O.	B.	O.	B.	O.	B.	O.	B.	O.	B.	O.
1	62.1	62.2	42.8	47.7	47.1	47.1	35.2	35.5	40.5	41.0	64.2	64.1	48.7	**49.6**
2	59.8	60.5	42.2	49.1	42.6	42.2	35.5	35.5	46.7	47.9	47.2	71.7	45.7	**51.2**
3	63.4	63.8	44.2	46.4	47.1	47.6	39.0	39.1	51.3	51.9	55.9	69.5	50.1	**53.0**
4	65.4	65.8	44.9	46.8	51.3	51.4	41.3	40.7	51.7	52.2	52.9	65.3	51.3	**53.7**
5	67.7	68.1	46.0	47.5	53.7	53.2	42.7	42.8	52.9	53.3	46.0	68.5	51.5	**55.5**
6	70.8	71.2	47.3	48.2	55.5	55.3	43.8	43.5	57.8	57.5	63.3	69.6	56.4	**57.6**

Table 19: COREFERENCE RESOLUTION results of multi-task models.

A.2 CHAPTER 4

The mean of coreference MUC, $B^3$and $CEAF_{\phi_4}F_1$ scores for the supervised baseline (B.) and coherence fine-tuned (O.) models are shown in Tables 18 and 19 respectively. (1) indicates LSTM + CNN, (2) indicates BERT-Tiny, (3) indicates BERT-Mini, (3) indicates BERT-Small, (4) indicates BERT-Medium, and (5) indicates BERT-Base encoders. ON, PC, PD-G, PD-W, WC, and WB stand for OntoNotes, PreCo, Phrase Detectives (Gutenberg), Phrase Detectives (Wikipedia), WikiCoref, and WinoBias respectively.

The macro-averaged F_1 score for the SRL supervised baseline (B.) and coherence fine-tuned (O.) models are shown in Tables 20 and 21 respectively. ON, C-WSJ, C-B, and EWT stand for OntoNotes, CONLL05-Wall Street Journal, CONLL05-Brown, and English Web Treebank respectively.

A.3 CHAPTER 6

A.3.1 *Implementation and Reproducibility Details*

Following Rothe et al. (2020), the encoder and decoder of RobertaS2S and RobFame models are initialized with public RoBERTa checkpoints. The encoder and decoder parameters are shared in both cases. Only the encoder-decoder attention parameters are initialized randomly. For RobFame, the focus attention parameters are also randomly ini-

	ON		C-WSJ		C-B		EWT		Average	
	B.	O.	B.	O.	B.	O.	B.	O.	B.	O.
1	72.14	72.14	67.33	67.72	63.41	63.26	67.64	67.84	67.63	**67.74**
2	65.05	65.01	53.02	52.91	51.62	52.56	57.79	57.83	56.87	**57.08**
3	77.32	77.34	68.74	68.59	65.75	65.46	70.22	70.53	**70.51**	70.48
4	81.21	81.21	73.11	73.45	69.80	70.42	72.91	72.85	74.26	**74.48**
5	82.30	82.32	75.24	75.21	70.21	69.96	74.73	74.79	**75.62**	75.57
6	85.88	85.97	78.93	78.83	75.01	75.30	78.00	77.99	79.46	**79.52**

Table 20: SEMANTIC ROLE LABELING results of single-task models.

	ON		C-WSJ		C-B		EWT		Average	
	B.	O.	B.	O.	B.	O.	B.	O.	B.	O.
1	72.92	72.58	68.02	67.82	60.98	60.52	67.21	67.27	**67.28**	67.05
2	63.91	64.03	52.16	52.19	52.60	52.97	57.94	58.23	56.65	**56.85**
3	77.68	77.69	66.10	66.19	69.83	69.86	70.78	70.77	71.10	**71.13**
4	81.08	81.10	69.89	70.07	73.45	73.67	74.48	74.25	74.72	**74.77**
5	84.45	84.47	73.05	73.35	77.52	77.68	76.55	76.55	77.89	**78.01**
6	86.41	86.40	76.59	76.47	79.34	79.30	78.67	78.58	**80.25**	80.19

Table 21: SEMANTIC ROLE LABELING results of multi-task models.

tialized. We experiment with large RoBERTa checkpoints with 24 layers, a hidden size of 1024, filter size of 4096, 16 attention heads, and a vocabulary with 50K sentence pieces (Kudo and Richardson, 2018). RobertaS2S has around 455M parameters and RobFame has 463M parameters, with an additional 8M parameters. Our Pegasus and PegFame implementation also have the same configuration, except for the encoder-decoder attention parameters which are pretrained.

We used Cloud TPU v3 accelerators for training. All models are fine-tuned on the target task using Adam with a learning rate of 0.05. We use a linear learning rate warm up with 40k steps, normalized by the square root of the hidden size, and a square root decay. We do not perform any tuning on these hyperparameters. We use a global batch size of 128 document-summary pairs. We adapt to different number of training steps depending on the training data sizes. Models are trained for 400k and 200k steps for CNN/DM and XSUM respectively, saving check-points every 1000 steps. We choose the best model based on ROUGE-L performance on the respective validation set.

The vocabulary for functional tokens F is constructed by taking the most frequent sentence pieces in the training set. We tune $|F|$ using the respective validation sets; for XSUM, we choose $f = 500$ frequent sentence pieces and for CNN/DM, $f = 1000$. For all our experiments with the FAME models, the beam size is set to 4.

We use Cloud TPU v3 accelerators for computing entailment scores which takes about 20 minutes for the two datasets' test sets. Question

generation and answering for Feqa are run on a NVIDIA V100 GPU, and it takes between 8-12 hours for one setting of each test set.

Models	CNN/DM		
	R1	R2	RL
Lead	39.60	17.70	36.20
PtGen (See et al., 2017)	39.53	17.28	36.38
Bottom-Up (Gehrmann et al., 2018)	41.22	18.68	38.34
SAGCopy (Xu et al., 2020)	42.53	19.92	39.44
MASS (Song et al., 2019)	42.12	19.50	39.01
UniLM (Dong et al., 2019a)	43.33	20.21	40.51
BART (Lewis et al., 2019)	44.16	21.28	40.90
T5 (Raffel et al., 2019a)	43.52	**21.55**	40.69
PEGASUS (C4, Zhang et al., 2019)	43.90	21.20	40.76
PEGASUS (HugeNews, Zhang et al., 2019)	44.17	21.47	41.11
ProphetNet (Qi et al., 2020)	**44.20**	21.17	**41.30**
RobertaS2S (Rothe et al., 2020)	39.88	18.66	37.22
RobFame (ours)	40.27	18.43	37.51
PEGASUS (ours)	42.62	20.38	39.61
PegFame (ours)	**42.95**	**20.79**	**39.90**

Table 22: Abstractive summarization results on CNN/DM datasets. The **underlined bold** results are from the best performing models from literature and the **bold** results are the best performing FAME models.

A.3.2 *Abstractive Summarization Results on CNN/DailyMail*

The CNN/DM dataset (Hermann et al., 2015) consists of 287,227/13,368/11,490 training/validation/test document-summary pairs. The CNN/DM summaries are in the form of bullet-point story highlights and exhibit a high degree of extraction, requiring the models to learn to copy from the source documents. The XSUM summaries, on the other hand, are extreme, in that the documents are summarized into single-sentence summaries with a high level of abstractiveness. For comparison, the XSUM summaries show a much larger percentages of novel constructions than found in CNN/DM summaries (35.8/83.5/95.5/98.5 vs. 16.8/54.3/72.4/80.4 novel 1/2/3/4-grams). We use the original cased version. During training, the input documents are truncated to 512 tokens and the length of the summaries are limited to 128 tokens.

Table 22 and 23 present complete results for CNN/DM dataset. We see similar kind of improvements as observed in Table 11, except for ROUGE-2 for RobFame which is 0.23 points worse than the RobertaS2S baseline. Our best model PegFame performs better than both copy mechanism models: LSTM-based PtGen (See et al., 2017) and Transformer-based SAGCopy (Xu et al., 2020). PegFame performs worse when compared with T5 (Raffel et al., 2019a), the original PE-

Models	Len.	Rep. %	R1(P%) With doc.	doc. → sum. ent. (↑)	doc. → sum. ¬ cont.	Feqa acc.	Feqa avg.(#Q)	BERTSc.
RobertaS2S	52.1	77.6	92.7	88.8	96.4	37.3	18.1	76.0
RobFame	55.5	79.6	92.5	87.3	96.3	35.2	19.3	76.1
Pegasus	58.1	69.4	95.0	90.9	97.5	40.3	21.0	76.8
PegFame	58.5	71.0	95.3	**91.0**	**97.6**	**41.1**	21.1	**76.9**

Table 23: Faithfulness and qualitative assessment of summaries on Cnn/Dm dataset.

gasus (Zhang et al., 2019c) and ProphetNet (Qi et al., 2020). This can be expected as the number of parameters in PegFame is almost half of T5 or ProphetNet, and is 100M less than that in the original Pegasus.

RobFame performs worse than RobertaS2S on both ent. and Feqa measures for Cnn/Dm, similar to rouge-2 in Table 22. We hypothesize that this is due to the extractive nature of the Cnn/Dm dataset and the fact that it is not able to copy tokens from the input to the necessary extent as the encoder-decoder attention is not pre-trained. Moreover, Feqa scores for RobertaS2S and RobFame may not be fully comparable due to variation in their summary lengths and the number of Feqa questions generated; the RobFame summaries, on average, are 3 words longer and generate 1.2 more questions than that of RobertaS2S. Nevertheless, we don't see this kind of drop in ¬cont. scores (i.e., summary not contradicting, either entailed by or neutral to the document) and BERTScores.

A.3.3 *Text Editing Results*

We also train the Fame models on two text editing tasks: (i) for sentence fusion – the problem of combining multiple sentences into a single coherent sentence – we used the "balanced Wikipedia" portion of the DiscoFuse dataset Geva et al. (2019), and (ii) for split-and-rephrase – the reverse task of sentence fusion – we used the WikiSplit dataset Botha et al. (2018), which consists of 1M examples of sentence splits extracted from the Wikipedia edit history. As the name suggests, both text editing tasks require a low degree of abstraction.

For both the tasks, we train the models for 300k steps with a global batch size of 256. The input and output are padded to a length of 128, which covers 100% of the training, evaluation and test data. The vocabulary for functional tokens F is constructed by taking the top 100 and 500 sentence pieces for DiscoFuse and WikiSplit respectively.

We report corpus-level BLEU[2], the exact match accuracy, and SARI scores Xu et al. (2016)[3]. The results can be seen in Table 24. The vanilla Pegasus model already beats the current state-of-the-art on both Dis-

2 We use NLTK v3.2.2 with case sensitive scoring to estimate BLEU scores.

3 SARI is a lexical similarity metric which compares the model's output to multiple references and the input in order to assess the model's ability to add, delete, and keep an n-gram. It's implementation is available at: https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/utils/sari_hook.py.

DiscoFuse	Exact	SARI	BLEU
(Geva et al., 2019)	51.1	84.5	–
LaserTagger (Malmi et al., 2019)	53.8	85.5	–
Felix (Mallinson et al., 2020)	61.3	88.8	–
RoBERTaS2S (Rothe et al., 2020)	66.6	90.3	–
PEGASUS (ours)	67.4	90.5	95.8
PEGFAME (ours)	**67.8**	**90.7**	**95.9**
WikiSplit	Exact	SARI	BLEU
(Botha et al., 2018)	14.3	61.5	76.4
LaseTagger (Malmi et al., 2019)	15.2	61.7	76.3
RoBERTaS2S (Rothe et al., 2020)	16.4	63.8	**77.4**
PEGASUS (ours)	16.6	**64.1**	**77.4**
PEGFAME (ours)	**16.8**	**64.1**	77.3

Table 24: Text editing results on Discofuse and WikiSplit. The underlined scores beat the current state-of-the-art and the **bold** scores are the new state-of-the-art.

coFuse and WikiSplit. The PEGFAME model performs better, albeit by a small margin, on all metrics on DiscoFuse. On WikiSplit, it has a higher exact match accuracy while maintaining the SARI score and performs 0.1 BLEU worse than PEGASUS.

A.3.4 *Controlled Generation with focus attention using Top-k tokens*

Table 25 presents results from our controlled summary generation experiments with top-k tokens from t_X using focus attention ($\text{Focus}_{\text{top,k}}$) on the XSUM test set. In Figures 20 and 21, we describe how ROBFAME consistently outperforms PEGFAME at lower values of $k \in \{50, 100, 200, 500, 1000\}$ due to their peaky and smooth t_X, respectively. While Figure 21 only plots ROUGE-1 F1 scores, Table 25 additionally reports ROUGE-2, ROUGE-L, entailment, Feqa, and BERTScores. Figure 27 presents predictions from models using $\text{Focus}_{\text{top,k}}$ for the article presented in Figures 18 and 26.

A.3.5 *Diverse Summarization with* $\text{Div}_{\text{top,k}}$, $\text{Div}_{\text{nucleus}}$ *and* $\text{Focus}_{\text{sample,k}}$

Figures 28, 29, 30, 31, 32, 33, and 34 show the diverse summaries generated using $\text{Focus}_{\text{sample,k}}$, $\text{Div}_{\text{top,k}}$ and $\text{Div}_{\text{nucleus}}$ sampling methods for the article shown in Figure 26.

A.4 CHAPTER 7

A.4.1 *Language Partitions*

The languages from the following families in UD are held out for evaluation (16 treebanks, 14 languages in total): Northwest Caucasian

Metrics	ROUGE R1	ROUGE R2	ROUGE RL	ent.	Feqa	BERTScore
RobertaS2S	41.45	18.79	33.90	39.1	19.8	80.6
RobFame	**42.15**	**19.68**	**34.81**	**41.3**	**21.2**	**80.8**
RobFame ($Focus_{top,k=50}$)	30.90	10.60	24.85	27.1	10.6	74.2
RobFame ($Focus_{top,k=100}$)	33.62	12.39	27.14	30.3	12.4	74.2
RobFame ($Focus_{top,k=200}$)	35.99	14.12	29.23	32.4	13.9	77.3
RobFame ($Focus_{top,k=500}$)	38.29	16.04	31.30	35.8	15.9	78.6
RobFame ($Focus_{top,k=1000}$)	39.58	17.18	32.49	37.3	17.3	79.3
RobFame ($Focus_{top,k=10000}$)	41.58	19.13	34.30	40.7	20.2	80.5
Pegasus	44.85	22.26	37.03	43.6	24.5	81.7
PegFame	**45.31**	**22.75**	**37.46**	**44.8**	**24.8**	**81.9**
PegFame ($Focus_{top,k=50}$)	24.30	7.52	19.32	20.8	8.0	68.8
PegFame ($Focus_{top,k=100}$)	27.77	9.26	22.09	24.1	9.3	71.3
PegFame ($Focus_{top,k=200}$)	31.05	11.14	24.82	27.0	10.8	73.6
PegFame ($Focus_{top,k=500}$)	34.99	13.65	28.19	31.0	13.0	76.2
PegFame ($Focus_{top,k=1000}$)	37.40	15.30	30.16	33.6	14.9	75.9
PegFame ($Focus_{top,k=10000}$)	42.76	19.89	34.97	40.2	20.1	80.5

Table 25: Assessment of controlled summary generation with focus sampling $Focus_{top,k}$ on the XSum test set. We experiment with limiting Fame models to different sizes of vocabulary V_k using the topic distribution t_X; in particular, we experiment with $k = \{50, 100, 200, 500, 1000, 10000\}$. We also report numbers for RobertaS2S, RobFame, Pegasus and PegFame, using the whole vocabulary of size 50k. The **bold** results in each block are the best performing RobertaS2S-based and Pegasus-based models.

(Abaza), Mande (Bambara), Mongolic (Buryat), Basque, Tupian (Mbya Guarani), Creole (Naija), Tai–Kadai (Thai), Pama–Nyungan (Warlpiri), Austronesian (Indonesian, Tagalog), Dravidian (Tamil, Telugu), Niger-Congo (Wolof, Yoruba). As all 8 languages in TiDiQA belong to families with at most 2 members in the dataset, we randomly create two partitions: in the former, Finnish, Korean, Bengali, and Arabic are used for evaluation, and the others for training; in the latter, Russian, Indonesian, Telugu, and Swahili are used for evaluation, and the others for training.

A.4.2 *Hyperparameter Setting*

POS TAGGING. For POS tagging: (i) the batch size was 32, (ii) the maximum sequence length was 128, (iii) the number of epochs was 20, with a patience limit of 10, (iv) both outer and inner learning rates were 5×10^{-5}, (v) the number of episodes per iteration was 32, (vi) the number of inner loops per outer update was 4, (vii) the number of shots (k) during training was 30, and (viii) the hidden layer dropout probability for the classifier was 0.2.

Gold	Australia has expelled an Israeli diplomat saying Israel was behind the forging of Australian passports linked to the murder of a Hamas operative in Dubai.
Article	Australia's foreign minister said these were "not the actions of a friend". The UK took similar action in March, after concluding that Israel was responsible for the use of forged UK passports in the plot. The Israeli foreign ministry said Australia's decision was disappointing. Ministry spokesman Yigal Palmor said it was "not in line with the importance and the quality of the relationship between our countries". 'Sorrow not anger' At least four forged Australian passports were used in the killing of Mahmoud al-Mabhouh in Dubai in January. The originals belonged to Australians living in Israel. The Australian government said a police investigation had left it in no doubt that the Israeli authorities were behind "the abuse and counterfeiting of the passports". As a result Foreign Minister Stephen Smith asked Israel to withdraw a diplomat, whom he did not identify. "The decision to ask Israel to remove from Australia one of its officers at the Israeli embassy in Canberra is not something which fills the Australian government with any joy," he said. "On the contrary, the decision was made much more in sorrow than in anger." Passports from France, Ireland, Germany and Britain were used in the operation, and in March, the British government expelled an Israeli diplomat from London. The Israeli government has said there is no proof that it was behind the killing, although Dubai officials have said they are 99.9% sure that agents from Mossad were responsible.
RobertaS2S	Australia has asked Australia to withdraw an Israeli diplomat from its embassy in Canberra after an alleged plot to kill a Abu Dhabi militant in Dubai.
RobFame	Australia has asked Israel to withdraw one of its diplomats from its embassy in Canberra after it admitted it used forged passports.
Pegasus	Australia has expelled an Israeli diplomat after concluding that forged Australian passports used in the killing of a Hamas militant in Dubai were issued by Israel.
PegFame	The Australian government has expelled an Israeli diplomat over the use of forged Australian passports in the killing of a Hamas militant in Dubai.

Figure 26: A 2010 BBC article from the XSum testset, its human written summary and model predictions from RobertaS2S, and Pegasus, with and without Fame. The text in orange is not supported by the input article.

QA. (i) the batch size and k were reduced to 12 due to memory constraints, (ii) the maximum context length was 336, and the document stride was 128, (iii) the maximum question length was 64, (iv) the inner and outer learning rates were 3×10^{-5}.

For all J baselines, we used a uniform language sampler, since proportional sampling performed worse. As an optimiser, we chose Adam with a learning rate of 5×10^{-5}, a weight decay of 0.1; we

	RobFame
(Focus$_{top,k=50}$)	Australia has said it will not be expelled an ambassador from Australia following the alleged s agent for the so-called Arab Arab State.
(Focus$_{top,k=100}$)	Australia has said it will not be expelled an ambassador from Australia following the killing of a terror agent in the Arab world.
(Focus$_{top,k=200}$)	Australia has said it will not be expelled an ambassador from Australia following the killing of an Australian terror suspect in the Arab world.
(Focus$_{top,k=500}$)	Australia has asked Israel to end its diplomatic investigation into an alleged plot to murder an Australian terror suspect.
(Focus$_{top,k=1000}$)	Australia has asked Israel to strip an ambassador from its embassy following the death of an Arab man in Dubai.
(Focus$_{top,k=10000}$)	Australia has asked Israel to withdraw one of its diplomats from its embassy in Canberra following the death of a terror suspect.
	PegFame
(Focus$_{top,k=50}$)	The Israeli government has been expelled from the country after it was found that the country's security agency, the Israeli intelligence agency, was to be to be found to have used a number of the country's out-of-country p when it was used in the Emirates car-j best.
(Focus$_{top,k=100}$)	The Israeli government has been expelled from the country after it was found that the country's security agency, the Israeli intelligence agency, had used the country's visas in the Emirates terror.
(Focus$_{top,k=200}$)	The Australian government has expelled an Israeli diplomats after it found that the country's security agency, the Israeli intelligence agency, had used the country's visas in the Emirates terror attack.
(Focus$_{top,k=500}$)	The Australian government has expelled an Israeli diplomatic staff after accusing the country's security agency, the Israeli intelligence agency, of using a number of Australian visas in the Emirates terror attack.
(Focus$_{top,k=1000}$)	Australia has expelled an Israeli diplomatic staff after accusing the country's security agency, the Israeli military's intelligence agency, of being responsible for the use of Australian visas used in the killing of a Palestinian.
(Focus$_{top,k=10000}$)	Australia has expelled an Israeli diplomat over the use of forged Australian passports in the killing of a Hamas militant in Dubai.

Figure 27: Model predictions with focus sampling Focus$_{top,k}$, a controlled generation setting. The text in orange is not supported by the input article. We note that with smaller values of k, both RobertaS2S-based and Pegasus-based models tend to hallucinate more often.

clipped the gradient to a maximum norm of 5.0. For all MAML models, we performed 4 updates in the inner loop, both during training and fast adaptation (few-shot learning). We ran our experiments on a 48GB NVIDIA Quadro RTX 8000 GPU with Turing micro-architecture. Each run took approximately 2 hours for training and 3 hours for few-shot learning and evaluation.

Dataset	k	J	B	MM	NP	MM+	NP+
abq_atb	0	14.34	24.11	20.41	26.81	16.42	22.55
	5	33.32±5.58	35.03±5.85	37.61±4.57	39.23±5.41	37.41±7.3	39.95±5.93
	10	37.52±2.28	40.38±5.75	43±3.63	42.7±5.3	43.57±7.31	46.56±4.91
	20	40.83±6.53	44.92±7.08	45.83±5.59	45.25±7.26	45.21±9.4	48.12±7.19
bm_crb	0	29.56	30.85	29.2	28.57	30.44	30.22
	5	45.6±3.47	50.83±3.33	46.04±3.95	45.14±3.73	48.2±3.74	48.32±3.63
	10	49.75±1.23	54.4±2.73	50.35±2.7	50.01±2.74	51.65±2.87	51.28±3.4
	20	54.03±1.52	57.53±1.68	53.12±2.16	53.39±1.85	54.38±1.85	53.89±2.08
bxr_bdt	0	48.85	51.71	50.41	50.49	54.21	51.94
	5	51.29±1.67	51.81±2.18	51.57±2.21	51.62±2.17	53.09±2.08	51.83±2.31
	10	53.64±0.96	54.95±1.68	54.18±1.53	54.25±1.63	55.47±1.8	55.17±1.43
	20	56.18±1.13	57.23±1.17	56.48±1.49	56.97±1.2	58.19±1.38	57.29±1.12
eu_bdt	0	70.2	71.76	73.22	72.57	73.54	73.29
	5	74.7±1.39	75.74±1.69	75.42±1.59	75.77±1.94	76.58±1.64	76.52±1.64
	10	76.51±2.38	78.1±1.25	77.52±1.01	78.08±1.21	78.73±1.36	78.19±1.38
	20	78.52±0.67	80.09±0.87	79.47±0.84	80.01±0.76	80.69±0.91	80.24±0.78
gun_thomas	0	32.06	35.72	33.91	31.97	33.87	33.84
	5	40.65±2.27	42.62±3.05	43.28±2.64	42.32±2.45	43.12±2.63	42.46±2.37
	10	44.06±0.99	45.65±2.33	45.92±2.59	45.23±2.31	46.98±2.49	45.41±2.25
	20	46.46±2.07	47.96±2.11	50.34±2.3	48.15±2.09	50.67±2.15	48.44±1.74
id_gsd	0	77.24	77.97	77.68	74.79	77.85	76.15
	5	82.2±1.22	83.47±1.22	82.72±1.47	82.35±1.68	83±1.5	82.47±1.57
	10	83.63±0.93	84.69±0.91	84.28±1.17	84.06±1.09	84.4±1.03	84.69±0.96
	20	84.75±0.61	85.75±0.59	85.82±0.58	85.35±0.68	85.94±0.66	85.86±0.69
id_pud	0	68.46	69.41	69.27	68.67	69.41	68.72
	5	73.07±1.39	73.96±1.5	73.5±1.48	74.17±1.43	74.52±1.46	73.82±1.56
	10	74.91±1.33	75.7±1.19	75.5±1.08	75.87±1.15	76.42±0.8	75.85±0.94
	20	76.17±0.57	77.18±0.72	77.06±0.49	77.28±0.68	77.75±0.57	77.39±0.71
pcm_nsc	0	61.97	40.78	45.77	40.76	41.21	56.83
	5	78.17±1.58	77.87±1.27	77.42±1.67	76.48±1.74	77.33±1.55	77.71±1.78
	10	80.06±1.24	79.28±1.25	78.96±1.1	78.41±1.1	78.71±0.94	80.03±1.37
	20	81.61±0.85	80.6±0.81	80.17±0.8	79.97±1	80.13±0.72	81.99±0.99

Table 26: POS tagging results on all evaluation languages: Part 1.

Dataset	k	J	B	MM	NP	MM+	NP+
ta_ttb	0	55.65	56.31	58.12	58.47	60.18	55.93
	5	72.29±2.03	72.39±2.21	71.37±1.7	72.28±2.46	72.34±2.13	70.19±2.3
	10	74.73±2.27	75.36±1.47	73.7±1.36	75.51±1.54	75.11±1.47	73.69±1.73
	20	76.23±1.19	77.56±1.38	75.75±1.39	77.83±1.33	77.44±1.3	76.29±1.49
te_mtg	0	75.21	75.87	77.49	75.43	76.28	76.29
	5	76.45±2.57	73.9±3.87	75.32±2.9	74.74±3.63	74.97±2.87	74.37±3.46
	10	78.68±1.74	77.16±2.55	78.26±2.09	77.55±2.29	77.57±2.12	76.94±2.83
	20	80.13±1.97	79.66±1.64	79.99±2.15	80±2.22	80.09±1.98	80.08±1.99
th_pud	0	42.51	42.71	43.76	43.3	46.81	43.07
	5	58.05±2.53	59.83±2.35	60.02±2.62	61.18±2.74	61.12±2.95	60.15±2.05
	10	61.71±2.17	63.57±1.72	63.85±1.9	65.14±1.67	65.4±1.87	63.34±1.75
	20	65.05±1.28	66.39±1.38	66.62±1.08	67.99±1.41	68.72±1.28	66.27±1.36
tl_trg	0	76.9	77.43	77.59	85.12	82.27	80.62
	5	83.01±3.52	82.95±3.66	84.09±4.75	84.5±4.14	84.4±4.01	84.01±4.64
	10	85.78±1.66	85.4±2.06	86.86±2.3	87.27±2.62	87.23±2.87	87.42±2.12
	20	87.27±2.04	87.48±2.32	88.69±1.96	89.1±2.34	89.2±1.85	89.24±1.86
tl_ugnayan	0	60.37	64.38	63.58	63.01	64.41	64.76
	5	74.8±1.86	76.35±2.27	75.73±2.01	75.2±2.37	78.13±2.01	76.91±2.44
	10	77.02±3.68	79.31±1.48	78.35±1.64	78.93±1.3	80.69±1.71	79.28±1.62
	20	78.91±1.44	82±1.07	80.86±1.04	81.14±1.32	82.66±1	81.71±1.31
wbp_ufal	0	26.64	24.55	28.62	27.21	27.96	30.18
	5	58±4.23	56.83±4.94	57.07±4.67	58.52±4.98	59.13±4.86	59.68±6.1
	10	64.72±1.72	63.34±4.41	64.51±3.43	65.94±3.88	65.2±4.03	66.32±3.63
	20	71.84±3.39	66.67±3.67	70.45±3.46	70.6±3.29	67.98±3.77	70.75±3.55
wo_wtb	0	34.79	33.05	34.72	34.11	34.09	35.27
	5	46.12±2.41	45.47±2.7	45.86±2.36	46.69±2.23	47.49±2.66	46.49±2.3
	10	50.01±2.03	48.49±1.69	49.13±2.18	49.69±1.79	50.97±2.1	49.67±2.1
	20	53.32±1.19	51.27±1.39	52.73±1.65	52.79±1.15	53.97±1.45	52.58±1.55
yo_ytb	0	41.46	47.34	45.31	45.59	50.45	49.1
	5	59.59±3.02	62.93±2.71	61.66±2.54	61.26±2.8	64.5±2.51	63.3±3.09
	10	63.34±	66.71±1.63	65.68±2	65.39±2.17	68.18±1.63	67.31±1.5
	20	67.23±1.06	69.14±1.19	69.45±1.01	68.56±1.43	70.9±1.1	69.58±1.25

Table 27: POS tagging results on all evaluation languages: Part 2.

RobFame (Focus$_{sample,k}$)

Australia has asked Israel to strip one of its diplomats from its embassy following the death of an Arab man in Dubai.

Australia has asked Israel to end its diplomatic investigation into an alleged plot to murder an Australian terror suspect.

Australia has asked Israel to strip one of its diplomats from its embassy in Australia over the death of a terror suspect.

PegFame (Focus$_{sample,k}$)

The Australian government has expelled an Israeli diplomatic staff after accusing it of using a number of Australian visas in the killing of a Palestinian in a car bombing.

The Australian government has expelled an Israeli diplomatic staff after it said the country was responsible for the use of Australian visas used in the killing of a Palestinian in a car bombing.

Australia has expelled an Israeli diplomatic staff after accusing the country's security agency, the Israeli military's intelligence agency, of being responsible for the use of Australian visas used in the killing of a Palestinian.

Australia has expelled an Israeli diplomatic mission after accusing the country's security agency, the Israeli military's intelligence agency, of being responsible for the use of Australian visas used in the killing of a Palestinian in the Arab city of Emirates.

The Australian government has expelled an Israeli diplomatic staff after it said the country was responsible for the use of Australian visas used in the killing of a Palestinian in the Middle East.

Figure 28: FAME model predictions with Focus$_{sample,k}$ ($k = 10000$). The text in orange is not supported by the input article.

A.4.3 *Additional Experiments & Results*

ADDITIONAL RESULTS. Tables 26 and 27 contain POS tagging F_1 scores of all languages, for all models, in both zero and few-shot settings. Tables 28 and 29 show the exact match and F_1 scores for QA.

SINUSOIDAL REGRESSION. After delving into real-world, large-scale NLP applications, we additionally illustrate the effect of the alternative criteria on other ML domains. We run a proof-of-concept experiment on a toy task where we can fully control the distribution of the training and evaluation data, viz. regression of a sinusoidal function.

For this task, we follow the same experimental setting and hyper-parameters of Finn et al. (2017): combinations of amplitudes $a \in [0.1, 5]$ and phases $p \in [0, \pi]$ determine a set of tasks characterised by the function $y = \sin(x - p) \cdot a$. The inputs are sampled at random from the interval $x \in [-5, 5]$.

While both train and evaluation tasks in the original version were sampled uniformly from *identical* ranges, we also construct an alternative setting with *skewed* distributions sampled from disjoint ranges: during training, $a \in [2.5, 5]$ and $p \in [\frac{\pi}{2}, \pi]$; during evaluation, $a \in [0.1, 2.5]$ and $p \in [0, \frac{\pi}{2}]$.

For Minimax MAML, we aim at learning the distribution over tasks adversarially. In particular, we consider two separate discrete categori-

RobertaS2S ($\text{Div}_{\text{top,k}}$)

Australia has asked for an Ivan "shivers" officer to be asked to leave Australia after the performance of an Israeli flag was alleged to have been used as terrorism suspects in Dubai.

Australia has asked an Israeli ambassador to Sydney over an alleged implicated Australian diplomat alleging the murder of a Australian national in Dubai.

Israel has asked Israel to withdraw an Israeli ambassador from Canberra amid claims that the alleged invasion of its territory by a foreign agent was behind the murder of a terror suspect in Abuabad.

Australia has asked Israel to withdraw a diplomat Izzy Kanhuh, an Israeli diplomat involved in solving tensions over the sale of imported shotguns for the Dubai Abu Dhabiuddin bombing.

Australia has asked Australia to withdraw an ambassador from the country, amid a growing row over the alleged role of an Israel-based Abu Abu Malak director of agents.

Australia has asked Israel to replace its ambassador over a fatal stabbing in Sydney last week.

Australia has asked Israel to withdraw an Egyptian diplomat following the suicide of a suspected Abu Abu Mabhulas in the Australian capital, Canberra.

Australia has asked Australia for an official withdrawal from its embassy in Sydney after the death of a Palestinian diplomat in a Dublin diplomatic fanbase earlier this month.

Australia has asked Israel to withdraw an Israeli diplomat as part of a probe into the alleged involvement in the murder of a Abu Abuab militant.

Australia has asked an Israeli diplomat to be withdrawn from the country over the Diamondad bombing of a Abu Waduh as part of an investigation into its 2002 murders of a Abu Abu Baye bomber.

RobFame ($\text{Div}_{\text{top,k}}$)

Australia has played down claims its state ambassador was involved in finding out why the Mossad spy agent was behind the Rio stabbing.

Australia has asked Israel to withdraw one of its diplomats after it confessed the so-called Mossad agent agent had used a fake Melbourne funery.

Australia says it will withdraw an envoy after the Israel spy agent accused of involvement in the murder of an Arab smuggler was suspended.

Australia has asked Israel to expel one of its citizens after the country leaked the state agent that led later a deadly mafia murder in Dubai.

Australia has asked Israel to withdraw its consulate at Canberra because from its embassy after it claimed it used the Falcon fuelling plan for a suicide bomb.

Australia has asked Israel to withdraw its support for Europe's embassy for its arrest of an Edinburgh diplomat over the death of a heroin smuggling gang.

Australia has asked Israel to remove an ambassador from its embassy over the shooting dead of an Australian man on a Dubai delivery scheme.

Australia is to withdraw a diplomat from its embassy in Canberra over allegations it worked on the mastermind for an alleged spying plot for the Mossad operation.

Australiachas asked Israel to withdraw an anonymous diplomat from its embassy following investigation into the passage of a Falcon recruiting device.

Australia has asked the Israeli embassy to pull out of its alleged response to the murder of a British terror suspect, accusing it of responsibility.

Figure 29: Diverse summaries predicted using RobertaS2S and RobFame models with $\text{Div}_{\text{top,k}}$.

cal distributions for amplitudes $\text{softmax}(\boldsymbol{\tau}_u^{(a)})$ and phases $\text{softmax}(\boldsymbol{\tau}_u^{(p)})$

RobertaS2S ($\text{Div}_{\text{nucleus}}$**)**

Australia says man hasenzelled an Israeli envoy following the arrest of one of its diplomats in Dubai from the countries' deepest-running terrorism resistant group. badly documented.

Australia has asked for Israel to out retrieving an Israeli diplomat who was expelled from the country after Australia accused the FBI of involvement in a 2013 murder in rogue Myersad drug smuggling operation.

Australia has asked Israel to remove an envoy from its embassy in Sydney in an escalating row over the killing of a Yazad Bin Ab alcohol dealer in the United Arab Emirates.

Australia has asked Israel to clarify its response to a data breach cull from rendition with a relapse of a suspected Abu Abuabuded jihadist.

Australia has asked Australia to pull out of Israel after an Israeli diplomat was accused of having used sreleased Australian agent Abuadab in the murder of an Abu Dhabi carrier.

Australia Herb Allen has led Australia's ambassadorsaints over an investigation into what was allegedly led by one of its diplomats at Nessadab consultancy in Dubai.

Australia has urged Israel to withdraw an ambassador pshorze over alleged links to the murder of a Sydney binnington.

Australia has asked the Israeli ambassador to Australia over an inferno at a Sydney diplomatic consulate for a senior recruiter which printers had wanted a Willis bin Laden agent to be charged.

Australia has asked Australia for the withdrawal of an Israeli ambassador after an investigation into it was linked to a Vietnam-based gang in which a young dungeonsad spy was killed.

Australia has asked Israel for an emotional withdrawal from its embassy in Canberra, accusing an Israeli diplomat of involvement in a feuding plot to kill a terror suspect.

RobFame ($\text{Div}_{\text{nucleus}}$**)**

Australia has asked Israel to withdraw an Israeli official over a Team Mossad bomb plot that left one of its suspects in the Dubai Arab desert.

Australia has asked all Israeli diplomats to leave Canberra after the living place of an alleged Russian special forces agent was identified at the email bug held bymacadad.

Australia is to withdraw an official sensitivity inquiry from its foreign ministers after Israel was accused of involvement in a plot to kill a Dubai terror suspect.

Australia has asked Israel deep back into allegations it carried out a wanted plot Cunning deaths in a Dubai plot by Mossad agents.

AustraliaplayedAX has asked the Israeli government to withdraw an official language envoy from its embassies following the killing of a murdered cons consulate officer.

Australia has asked an Israeli official to withdraw an official ambassador after it made a murder in a deadly shooting Presumably by Mossad.

Australia has asked Israel over allegations that an agent used forged passports to plot the Woolstroken murder by agentsbased in Pakistan.

Australia has asked Israeli authorities to withdraw an official diplomat from Australia after the mafia was accused by the Israel embassy of contributing to its alleged failed murder of an Alquer Arab Shia terrorist.

Australia has asked an Israeli embassy to withdraw a diplomat from Australia following the Jewlands' murder of an unnamed man.

Australia has annexed its embassy up tolishes at the start of the year after Israel confirmed it assessed the role of an undercover officer during the Dubai heroinmer plot.

Figure 30: Diverse summaries predicted using RobertaS2S and RobFame models with $\text{Div}_{\text{nucleus}}$.

RobFame ($\text{Focus}_{\text{sample,k}}$, $\text{Div}_{\text{top,k}}$)

Australia has asked Israel to answer the decision to honour its state ambassador following the alleged involvement in the killing of a Dubai terror suspect.

Australia has asked Israel for a second diplomat to be expelled from Australia after an alleged plot to murder a man in a bomb plot linked to Mossad.

Australia has asked Israel to make a state diplomat its top diplomat after an alleged plot to bomb an Arab Emirates terror operation was blamed on a terror agent.

Australia has asked Egypt to end its diplomatic at-top diplomatic response to the murder of a top Arab diplomat in the Arab world.

Australia has asked Israel to be expelled from the embassy in Australia following the death of a Sydney spy in a spy investigation.

Australia has asked Israel to to strip an diplomat of its consulate from its embassy since a deadly operation against the Mossad spy agent at a terror squad in Australia last month.

Australia has asked the Israel embassy to withdrawing its diplomats following the death of an Arab man by Mossad agents.

Australia has asked Israel to end the original accusations that a diplomat is responsible for the killing of an agent from Mossad.

Australia has asked Israel to answer the investigation that admitted its diplomats used his agent as a suicide bomb in a Dubai plot.

Australia has asked Israel to support its ambassador after it admitted being involved in the murder of a suspect in the deadly one-off terror killing in a Melbourne bomb attack.

RobFame ($\text{Focus}_{\text{sample,k}}$, $\text{Div}_{\text{nucleus}}$)

Australia has asked Israel to expelled an embassy diplomat over a deadly Sydney plot to spy on the Mossad operation.

Australia has asked Israel to end its diplomatic inruru from Australia after it accused its diplomatic staff of involvement in last year's deadly attack on a Melbourne terror attack.

Israel has asked Israel to make an embassy ambassador over a deadly email killing of a man in a terror plot.

Australia has asked Israel to strip its diplomatic staff of its passport following an alleged plot to murder a Dubai terror suspect.

Israel has asked Israel to expelled one of its diplomats after the Mossad agent accused a Melbourne man of being the agent for the Mossad spy agent for his role in an alleged plot to murder a man.

Australia has asked Israel to strip a top envoy from his embassy following its investigation into the killing of an alleged spy in a Melbourne email plot.

Australia has asked Israel to expelled one diplomat following allegations it used a military agent to spy for Mossad.

Australia has asked the Israel embassy to be expelled from Australia after an Australian diplomat was found guilty of his role in the murder of an Australian terror suspect.

Australia is to expelled its top diplomat from Australia after his country was accused by the UN of being responsible for an alleged plot to murder a Melbourne-Arab m intelligence agent.

Australia has asked Israel to strip an ambassador from its embassy, in response to the death of a Sydney-from-agent for the so-called "Mossad, was responsible".

Figure 31: Diverse summaries predicted using RobertaS2S and RobFame models with $\text{Focus}_{\text{sample,k}}$.

over their respective ranges discretised into 1,000 atoms. Hence, the

probability of a task with the i-th amplitude value and the j-th phase value is simply $\tau_i^{(a)} \times \tau_j^{(p)}$.

The results for sinusoidal regression are shown in Figure 37. Vanilla MAML (Bayes criterion) consistently outperforms the minimax criterion when the task distribution is identical; on the other hand, the reverse occurs when the task distribution is skewed. MM performs much better in this case, with the gap in performance increasing as the shots k decrease. This verifies our hypothesis that the minimax criterion should benefit out-of-distribution regression tasks.

Dataset	k	J	B	MM	NP	MM+	NP+
Arabic	0	48.97	49.29	51.47	51.36	49.4	48.64
	5	52.2±3.92	50.19±3.52	53.38±3.52	51.48±3.2	49.27±3.89	51.27±4.75
	10	54.51±2.47	52.81±2.93	54.96±2.93	53.67±2.16	52.05±3.27	53.67±3.43
	20	56±1.85	54.64±1.86	56.59±1.56	55.43±1.86	54.45±1.94	55.78±2.13
Bengali	0	45.13	46.02	51.33	44.25	45.13	51.33
	5	46.32±3.48	45.3±3.11	50.76±3.03	47.22±3.3	45±2.98	49.45±3.17
	10	47.22±3.15	46.44±3.08	50.83±2.94	49.39±3.84	45.98±3.7	50.01±3.22
	20	49.47±3.54	48.24±4.15	52.37±3.57	50.21±3.62	47.68±3.31	51.24±3.03
Finnish	0	42.33	43.61	47.95	49.36	47.83	46.42
	5	46.5±4.96	45.75±3.21	47.69±3.48	48.75±3.21	45.66±3.53	47.57±4.21
	10	48.56±2.65	47.25±2.81	49.43±2.78	50.28±3.1	46.85±2.77	48.55±3.1
	20	49.81±2.09	48.82±2.77	50.43±2.34	52.22±3.01	48.18±2.48	50.89±2.49
Korean	0	50	50.72	53.62	48.55	51.45	53.62
	5	51.37±2.52	49.5±2.76	51.87±2.11	49.52±2.48	49.57±2.35	52.17±2
	10	52.63±2.41	50.63±2.46	52.29±1.85	50.1±2.29	50.29±2.51	53±1.93
	20	54.07±2.11	51.88±2.15	53.55±1.91	51.87±2.03	51.71±2.13	53.67±2.16
Indonesian	0	56.46	51.86	54.87	56.28	52.74	56.28
	5	57.99±2.94	55.49±3.18	56.04±2.99	57.61±2.7	55.53±3.82	55.39±2.67
	10	59.4±2.49	57.11±2.81	58.54±2.49	58.59±1.96	57.08±2.84	56.86±1.95
	20	60.99±2.09	58.99±2.51	60.76±2.21	59.9±1.69	59.11±2.07	57.95±1.94
Russian	0	44.21	43.23	41.01	40.39	41.01	37.44
	5	49.45±4.36	47.41±3.92	46.66±4.01	46.83±4.34	46.2±4.61	44.09±5.38
	10	51.84±3.04	49.66±2.83	48.72±3.56	48.81±3.79	47.97±4.43	47.66±4.05
	20	53.6±2.45	50.72±2.55	51.05±2.8	51.5±2.75	50.47±2.45	50.25±2.96
Swahili	0	43.49	45.29	41.88	41.48	45.69	45.29
	5	46.47±5.11	49.07±4.31	48.9±4.88	47.6±4.21	48.8±4.28	47.32±4.3
	10	50.06±4.13	51.37±3.45	51.1±3.83	50.37±3.72	49.79±3.59	49.96±3.88
	20	54.02±3.06	53.82±2.63	53.94±2.54	52.16±2.89	52.51±3.47	52.65±3.26
Telugu	0	43.5	42.75	44.54	42	41.7	45.14
	5	45.97±2.85	44.58±3.44	45.33±3.91	44.89±3.44	41.87±5.35	42.92±5.36
	10	48.11±3.4	46.64±3.1	47.59±2.95	46.4±2.69	45.32±4.23	46.3±3.51
	20	50.1±2.55	49.08±2.42	49.21±2.77	48.8±1.97	47.11±2.71	48.88±2.91

Table 28: QA exact-match results on all evaluation languages.

Dataset	k	J	B	MM	NP	MM+	NP+
Arabic	0	65.57	67.38	66.59	67.44	64.98	65.45
	5	68.4±3.82	67.09±3.51	69.66±3.45	67.59±3.26	65.92±3.93	67.76±4.86
	10	70.56±2.47	69.55±2.88	71.35±2.83	69.82±2.15	68.82±3.64	70.28±3.63
	20	72.14±1.78	71.14±1.87	73.21±1.46	71.55±1.88	71.5±1.99	72.26±2.31
Bengali	0	57.24	62.57	66.29	59.64	60.28	62.86
	5	59.27±2.79	60.04±2.97	64.65±2.84	61.71±3.1	59.46±2.72	61.85±2.65
	10	59.88±2.7	60.64±2.86	64.88±2.71	63.52±3.38	60.03±3.28	62.33±2.89
	20	62.11±3.15	62.1±3.51	65.93±3.07	64.31±2.95	61.72±2.96	63.86±2.72
Finnish	0	61.85	63.57	61.72	63.79	62.12	61.64
	5	61.48±3.47	61.76±2.63	61.66±2.84	62.66±2.8	60.49±2.42	61.57±3.94
	10	62.98±1.53	62.48±2.03	63.26±2.31	64.14±2.7	61.58±2.15	62.58±2.91
	20	63.81±1.57	63.66±2.07	64.65±2.2	65.64±2.84	63±2.24	64.9±2.27
Korean	0	60.26	62.71	62.4	58.68	61.2	64.35
	5	61.31±2.47	60.82±2.47	61.67±2.17	59.31±2.34	59.27±2.33	62.13±2.01
	10	62.52±2.26	62.02±2.1	61.86±2.05	60.03±2.15	59.86±2.51	62.92±1.91
	20	64.04±2.01	63.08±1.87	63.43±1.89	61.84±1.97	61.52±1.89	63.55±1.95
Indonesian	0	69.96	65.99	70.05	70.82	68.02	70.19
	5	71.4±2.81	69.22±3.28	70.61±2.74	71.18±2.17	69.95±3.4	69.37±2.58
	10	72.69±2.27	70.79±2.78	72.79±2.53	72.23±1.77	71.33±2.46	70.93±1.96
	20	74.11±1.79	72.49±2.57	74.65±2.11	73.52±1.45	73.11±1.8	71.96±1.94
Russian	0	65.93	64.15	64.47	63.2	64.13	61.08
	5	66.96±1.52	65.11±1.5	65.03±1.39	65.13±1.33	64.58±1.45	62.17±1.61
	10	67.86±1.15	66.21±1.28	65.84±1.65	65.89±1.33	65.53±1.46	63.63±1.67
	20	68.7±1.01	66.85±1.38	66.94±1.45	67.1±1.38	66.4±1.19	65.01±1.82
Swahili	0	60.01	62.63	59.84	58.74	64.13	62.48
	5	60.21±4.38	62.7±3.37	62.48±3.72	61.9±3.41	63.43±3.38	61.81±4.06
	10	62.62±2.66	64.36±2.31	63.79±3.19	63.6±3.02	63.77±2.88	63.78±3.06
	20	65.18±2.09	65.89±2	66.27±1.99	65.48±1.7	66.21±2.19	66.12±2.11
Telugu	0	52.43	51.1	53.1	52.83	51.93	53.96
	5	60.99±5.15	59.6±6.05	59.21±6.17	61.27±5.05	57.91±6.64	57.21±7.33
	10	63.99±3.92	62.92±4.19	62.85±3.62	62.63±4.98	61.92±5.1	61.74±5.18
	20	65.96±2.37	65.29±2.15	64.53±3.06	65.63±1.61	63.67±2.58	64.9±3.26

Table 29: QA F1 results on all evaluation languages.

Pegasus ($\text{Div}_{\text{top,k}}$**)**

Australia has expelled an Israeli diplomat over the use of forged Australian passports in the killing of Hamas detainee Mahmoud al-Mabhouh in Dubai.

Israel has summoned the Australian ambassador to complain after the Australian government said forged passports used in the killing of a Hamas operative in Dubai belonged to Netanyahu's foreign ministry.

The Australian government has ordered Israel to withdraw an officer over the use of forged Australian passports used by the 2013 murder of a Lebanese opposition figure in Dubai.

The Australian government has expelled an Israeli diplomat over allegations that fake Australian passports were used to kill a Lebanese militant in Egypt two years ago.

Australia has asked Israel to withdraw a diplomat over the use of forged Australian passports to kill a Hamas operative in January.

Australia has expelled an Israeli diplomat in a row over the authenticated use of forged Australian passports in last year's killing of a Hamas figure in Dubai.

Australia says it is expulsion an Israeli diplomat in protest over Israel's alleged role in the killing of a Hamas militant in Dubai.

Australia has recalled a diplomat from Israel after accusing Berlin of fabricating false passports used in the assassination of a Hamas operative in Dubai.

Israel has been asked to withdraw an official from Australia, accusing it of complicity in the falsification of Australian passports used in the killing of a Hamas operative in Dubai in January.

Israel has withdrawn one of its diplomats after Canberra said it concluded that Passport Bureau agents participated in an internal Mossad plot to kill a Hamas operative in Dubai.

PegFame ($\text{Div}_{\text{top,k}}$**)**

Australia has expelled an Israeli diplomat after it concluded somebody close to Israel's security agency, Mossad, owned forged passports which were used to abduct a Hamas rocket maker.

Australia has expelled an Israeli diplomat over allegations that its intelligence agency Mossad was behind the use of forged passports in the killing of a suspected Palestinian militant.

Australia has expelled an Israeli diplomat amid accusations Israel-run Mossad used forged Australian passports in the killing of a Hamas militant.

Australia has expelled an Israeli diplomat in a dispute over the use of stolen Australian passports for a hit in the Dubai killing of a Lebanese militant earlier this year.

An Israeli diplomat has been expelled from Australia after a Sydney police team concluded that agents from the country's security agency Mossad took part in the poisoning of Egypt's president.

The Australian government has expelled an Israeli diplomat, after it concluded that his desk was responsible for the issuance of forged Australian passports used in the killing of a Hamas militant.

Australia has recalled her envoy from Israel, after finding that an Israeli diplomat was responsible for the counterfeiting of passports used by the Unesco agency director who was killed in Dubai.

The Australian government has asked Israel to withdraw from its Embassy in Melbourne after accusing it of using forged Australian passports to fund the killing of a Palestinian militant.

The Australian government has asked Israel to withdraw its ambassador for failing to acknowledge its role in the use of forged Australian passports in the killing of a British businessman.

Australia has formally demanded the removal of anIsrael diplomat in response to a decision to accuse the Jewish organisation Mossad of use of forged Australian passports mentioned in a Dubai bombing plot.

Figure 32: Diverse summaries predicted using PEGASUS and PEGFAME models with $\text{Div}_{\text{top,k}}$.

Pegasus ($\text{Div}_{\text{nucleus}}$**)**

Israel hasracuse withdrawn an envoy after the Australian government said it concluded that Israeli agents used forged passports used to kill a Dubai Bendigo businessman.

Australia has demanded the withdrawal of an Israeli diplomat, saying his arrival in Canberra was only necessary to deal with a spillover from the killing of a Hamas militant in Dubai in January.

The Australian government has recalled an Israeli diplomat over accusation that fake Australian passports used 436 kilometres (300 miles) from Canberra in the death of a Hamas militant were stolen by Israeli agents.

The Australian government has recalled aLatis from Israel for having strong evidence that their embassy was used to counterfeite passports used in the killing of a bidder in Dubai.

Australia has expelled an Israeli diplomat in a row about the use429 Australian passportsocally used in the killing of an intends in Dubai. slew of passports were used cosmetics in the killing.

Australia is seeking to expel an Israeli envoyrolet over the use of forged Australian passports in the murder of a militant in Dubaiselection.

Australia has expelled an Israeli diplomat after saying it was "certain" eagle-eyed undercover agents were Rhys Shapiro and Glenn Clift, who used forged Australian passports to kill an Soros geneticist in Dubai in 2015.

Australia has removed the Israeli ambassador following a decision to conclude that forged Australian passports used in the death of a Palestinian Deals in the DesertDex were collaborated from Israel.

Australia has recalled an Israeli diplomat, accusing Tel Aviv of "engaging in a pattern of alarming behaviour", after concluding that forged Australian passports were used in the killing of a Hamas operative in Dubai.

Israel has expelled one of its diplomats because of allegations that it helped Isabel al-Mabhouh, a British-based PalestinianORED to be kill in January, by using forged Australian passports.

PegFame ($\text{Div}_{\text{nucleus}}$**)**

Australia has summoned Idair Kernatic, a Jerusalem consulate official, inv summoned after the extraction of a document touting the use of forged passports for a deadly bomb plot.

Australia has recalled a diplomat from Israel, claiming Israel stole the original identities of passports used to kill a Hamas operative.

Australia has asked Israel Fever to withdraw a diplomat after New Zealand said Israeli agents used the fake local passports used to identify a keyoine Killer.

The Australian government has asked Israel to withdraw a diplomat after claiming the Jewish terror group Mossad used forged Australian passports in a plot to murder a Dubai imam.

Australia has expelled an Israeli diplomat after confirming fake Australian passports were used to help the killing ofmagazine boss Mahmoud al-Mabhouh in Dubai.

Australia has withdrawn a military characteristic of Israel after alleging its officials were behind the use of stolen Australian passportsxiety in a Dubai cash-in-transit plot.

Australia has expel an Israeli diplomat over allegations that the country's Mossad spy was behind at least Jong-Bam's Becket murder.

Australia has withdrawn an Israeli diplomat halves its embassy in Canberra over accusations the country's security service, Mossad, was responsible for issuing forged Australian passports.

Australia has asked Israel to withdraw one of its diplomats from Canberra after finding that phony Australian passports were used to kill an Egyptian cleric.

Australia has asked Israel to withdraw a diplomat after it said Israel was behinduse parts of forged Australian passports used in the bombing of a kayaker in Dubai.

Figure 33: Diverse summaries predicted using PEGASUS and PEGFAME models with $\text{Div}_{\text{nucleus}}$.

PegFame ($\text{Focus}_{\text{sample,k}}$**,** $\text{Div}_{\text{top,k}}$**)**

Australia has expelled an Israeli because "its anti-espionage agents" used visas from other nations to issue a Palestinian agent's body £2.3m (£"2.3,1) car and land lift to Hezbollah in the killing of a senior Palestinian in the city:

The Australian government has ejected an Israeli at its embassy over the use of Australia's visas in the killing of a terror attack in the city of D'scale.

Australia has expelled an Israeli diplomats following an investigation into the use of the country's travel services as cards used in a terror attack.

Australia has expelled an Israeli embassy transport staff after a police investigation found the country's intelligence agency, the Israeli intelligence agency, was at responsible.

Australia has expelled an Israeli embassy gathering after it said the country was responsible for the use of the use of Australia's Australian emails in an emailed attack on a former Australian consulate in the Arab world.

Australia has expelled an Israeli diplomatic side after accusing it of using at first issued Australian DNA test cards to produce the Irish agent in the stepped-up Emirates bombing.

Australia has expelled Israeli diplomats after its foreign minister said the country had been to be responsible for the use of staged Australian emails by Israeli intelligence.

Australia has told Israel to withdraw a diplomatic mission from its country - after it said it was "in no Petroleum to Finish" the killing of a Palestinian in the killing of a terror the network by the Israeli security agency, enzymes.

Australia has demanded Israel withdraw a diplomats following the Israel Security Service's use of Israeli-issued Australian travel visas to help one of its agents commit a terror attack.

PegFame ($\text{Focus}_{\text{sample,k}}$**,** $\text{Div}_{\text{nucleus}}$**)**

Australia has expelled an Israeli government agent after accusing it of using the use of Australian travel visas to help Israel's intelligence agency, arrest a Palestinian in a drug operation.

Australia has expelled an Israeli pulled over the use of Australian espionage proteins in a terror attack.

The Australian government says Israel should withdraw a senior police mission from its embassy following an investigation into the use of Australian gel-making equipment in the killing of a Palestinian in a car bombing in a Emirates airport.

Australia has expelled an Israeli diplomats for its support for a Palestinian that was used to hack the email messages of the former head of the intelligence agency, reasoning that the expulsion was "in the best security" of the two countries.

Australia has expelled Israel's second in service special operations, after accusing the country's intelligence agency, theahl, of a "poisoning".

Australia has expelled an Israeli government in protest at "the use" of an Australian denied diplomatic entry in a diplomatic killing in the Arab city of controversives.

Australia has expelled an Israeli posting at its embassy in a "diplomatic action", after it was found that Israeli agents had used issued Australian visas in the killing of an Egyptian man.

Australia has expelled an Israeli diplomatic, accusing it of using a home-grown Palestinian with a crime on the plane he was using to be arrested in the West.

The Australian government has expel an Israeli diplomatic team, following a public investigation into the use of Australian visas to help the killing of a Palestinian in the Emirates.

The Australian government has expelled an Israeli embassy consulate in Australia after saying it was "left in no suggestions" it was responsible for the use of Australian terror attack credentials in the killing of a Palestinian in the desert.

Figure 34: Diverse summaries predicted using PEGASUS and PEGFAME models with $\text{Focus}_{\text{sample,k}}$.

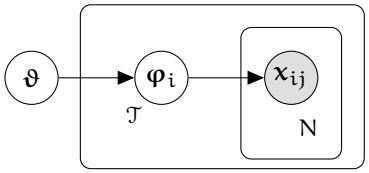


Figure 35: Bayesian graphical model of MAML, where the variable φ_i is parameterised as $\theta - \eta\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta, \mathcal{D}_{train})$.

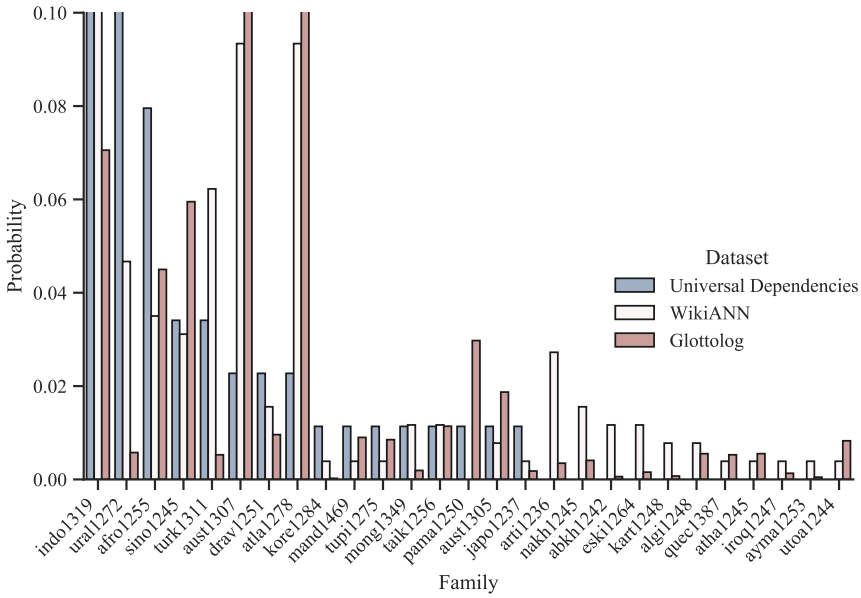


Figure 36: Empirical distribution of languages across families in 2 datasets (WikiANN and UD) and in the world, according to Glottolog. The families shown are a subset {(WikiANN ∪ Universal Dependencies) ∩ Glottolog}. The y-axis is truncated for the sake of clarity.

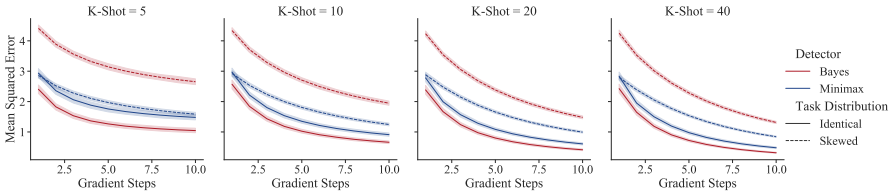


Figure 37: Mean Squared Error of MAML across gradient steps (from 1 to 10) of different criteria (B and MM) under identical and skewed task distributions. Each frame represents a separate run of fast adaptation with different amounts of target examples available (k-shot).

BIBLIOGRAPHY

Abdou, Mostafa, Cezar Sas, Rahul Aralikatte, Isabelle Augenstein, and Anders Søgaard (Nov. 2019). "X-WikiRE: A Large, Multilingual Resource for Relation Extraction as Machine Comprehension." In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, pp. 265–274. DOI: 10.18653/v1/D19-6130. URL: https://www.aclweb.org/anthology/D19-6130.

Abend, Omri and Ari Rappoport (Aug. 2013). "Universal Conceptual Cognitive Annotation (UCCA)." In: *ACL*. Sofia, Bulgaria, pp. 228–238. URL: https://aclweb.org/anthology/P13-1023.

Abend, Omri and Ari Rappoport (July 2017). "The State of the Art in Semantic Representation." In: *ACL*. Vancouver, Canada, pp. 77–89. URL: https://aclweb.org/anthology/P17-1008.

Ailem, Melissa, Bowen Zhang, and Fei Sha (2019). "Topic Augmented Generator for Abstractive Summarization." In: *CoRR* abs/1908.07026.

Anand, Pranav and Jim McCloskey (2015). "Annotating the Implicit Content of Sluices." In: *LAW@NAACL-HLT*.

Angeli, Gabor, Melvin Jose Johnson Premkumar, and Christopher D. Manning (July 2015). "Leveraging Linguistic Structure For Open Domain Information Extraction." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 344–354. DOI: 10.3115/v1/P15-1034. URL: https://www.aclweb.org/anthology/P15-1034.

Aralikatte, Rahul, Matthew Lamm, Daniel Hardt, and Anders Søgaard (2019a). "Ellipsis and Coreference Resolution as Question Answering." In: *CoRR* abs/1908.11141. arXiv: 1908.11141. URL: http://arxiv.org/abs/1908.11141.

Aralikatte, Rahul, Heather Lent, Ana Valeria Gonzalez, Daniel Herschcovich, Chen Qiu, Anders Sandholm, Michael Ringaard, and Anders Søgaard (Nov. 2019b). "Rewarding Coreference Resolvers for Being Consistent with World Knowledge." In: *EMNLP-IJCNLP*. Hong Kong, China, pp. 1229–1235. URL: https://aclweb.org/anthology/D19-1118.

Arnold, Sébastien M. R., Praateek Mahajan, Debajyoti Datta, Ian Bunner, and Konstantinos Saitas Zarkias (2020). "learn2learn: A Library for Meta-Learning Research." In: *arXiv preprint arXiv:2008.12284*. URL: https://arxiv.org/pdf/2008.12284.pdf.

Arumae, Kristjan and Fei Liu (2019). "Guiding Extractive Summarization with Question-Answering Rewards." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota, pp. 2566–2577.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives (2007). "DBpedia: A Nucleus for a Web of Open Data." In: *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*. ISWC'07/ASWC'07. Busan, Korea: Springer-Verlag, pp. 722–735. ISBN: 3-540-76297-3, 978-3-540-76297-3. URL: http://dl.acm.org/citation.cfm?id=1785162.1785216.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate." In: *CoRR* abs/1409.0473.

Balduzzi, David, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel (2018). "The Mechanics of *n*-Player Differentiable Games." In: *Proceedings of the 35th International Conference on Machine Learning*. Stockholm, Sweden, pp. 354–363. URL: http://proceedings.mlr.press/v80/balduzzi18a/balduzzi18a.pdf.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider (Aug. 2013). "Abstract Meaning Representation for Sembanking." In: *LAW*. Sofia, Bulgaria, pp. 178–186. URL: https://aclweb.org/anthology/W13-2322.

Banjade, Rajendra, Vasile Rus, and Nobal B. Niraula (2015). "Using an Implicit Method for Coreference Resolution and Ellipsis Handling in Automatic Student Answer Assessment." In: *FLAIRS*.

Bansal, Trapit, Rishikesh Jha, and Andrew McCallum (Dec. 2020). "Learning to Few-Shot Learn Across Diverse Natural Language Classification Tasks." In: *Proceedings of the 28th International Conference on Computational Linguistics*. online, pp. 5108–5123. DOI: 10.18653/v1/2020.coling-main.448. URL: https://www.aclweb.org/anthology/2020.coling-main.448.

Barzilay, Regina and Michael Elhadad (1997). "Using Lexical Chains for Text Summarization." In: *Intelligent Scalable Text Summarization*. URL: https://www.aclweb.org/anthology/W97-0703.

Beck, Amir and Marc Teboulle (2003). "Mirror descent and nonlinear projected subgradient methods for convex optimization." In: *Operations Research Letters* 31.3, pp. 167–175. URL: https://www.sciencedirect.com/science/article/pii/S0167637702002316.

Bender, Emily M. (Mar. 2009). "Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology." In: *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* Athens, Greece, pp. 26–32. URL: https://www.aclweb.org/anthology/W09-0106.

Bhatia, Parminder, Yangfeng Ji, and Jacob Eisenstein (Sept. 2015). "Better Document-level Sentiment Analysis from RST Discourse Parsing." In: *EMNLP*. Lisbon, Portugal, pp. 2212–2218. URL: https://aclweb.org/anthology/D15-1263.

Bickel, Peter J. and Kjell A. Doksum (2015). *Mathematical statistics: Basic ideas and selected topics, volume I*. CRC Press.

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. Springer.

Blackburn, Patrick and Johan Bos (2005). *Representation and Inference for Natural Language*. Stanford, CA: CSLI.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation." In: *The Journal of Machine Learning Research* 3, pp. 993–1022.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach (July 2020). "Language (Technology) is Power: A Critical Survey of "Bias" in NLP." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485. URL: https://www.aclweb.org/anthology/2020.acl-main.485.

Bordes, Antoine, Jason Weston, Ronan Collobert, and Yoshua Bengio (2011). "Learning Structured Embeddings of Knowledge Bases." In: *AAAI*.

Bos, Johan, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva (2017). "The Groningen meaning bank." In: *Handbook of linguistic annotation*. Springer, pp. 463–496.

Bos, Johan and Jennifer Spenader (Dec. 2011). "An Annotated Corpus for the Analysis of VP Ellipsis." In: *Lang. Resour. Eval.* 45.4, pp. 463–494. ISSN: 1574-020X. DOI: 10.1007/s10579-011-9142-3. URL: http://dx.doi.org/10.1007/s10579-011-9142-3.

Botha, Jan A., Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das (2018). "Learning To Split and Rephrase From Wikipedia Edit History." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 732–737.

Bower, GH and DG Morrow (1990). "Mental models in narrative comprehension." In: *Science* 247.4938, pp. 44–48. ISSN: 0036-8075. DOI: 10.1126/science.2403694. eprint: https://science.sciencemag.org/content/247/4938/44.full.pdf. URL: https://science.sciencemag.org/content/247/4938/44.

Brennan, Susan E., Marilyn W. Friedman, and Carl J. Pollard (1987). "A Centering Approach to Pronouns." In: *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*. ACL '87. Stanford, California: Association for Computational Linguistics, 155–162. DOI: 10.3115/981175.981197. URL: https://doi.org/10.3115/981175.981197.

Bunescu, Razvan (2003). "Associative Anaphora Resolution: A Web-Based Approach." In: *Proceedings of the 2003 EACL Workshop on The Computational Treatment of Anaphora*. URL: https://aclanthology.org/W03-2607.

Cai, Deng and Wai Lam (July 2020). "AMR Parsing via Graph-Sequence Iterative Inference." In: *ACL*. Online, pp. 1290–1301. URL: https://aclweb.org/anthology/2020.acl-main.119.

Cai, Jie and Michael Strube (2010). "Evaluation Metrics For End-to-End Coreference Resolution Systems." In: *Proceedings of the SIGDIAL 2010 Conference*. Association for Computational Linguistics,

pp. 28–36. URL: http://www.aclweb.org/anthology/W/W10/W10-4305.

Cao, Yue and Xiaojun Wan (Nov. 2020). "DivGAN: Towards Diverse Paraphrase Generation via Diversified Generative Adversarial Network." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2411–2421. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.218.

Cao, Ziqiang, Furu Wei, Wenjie Li, and Sujian Li (2017). *Faithful to the Original: Fact Aware Neural Abstractive Summarization*. arXiv: 1711.04434 [cs.IR].

Cardie, Claire and Kiri Wagstaff (1999). "Noun phrase coreference as clustering." In: *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Carlson, Greg (2006). "Anaphora." In: *Encyclopedia of Cognitive Science*. Wiley.

Carreras, Xavier and Lluís Màrquez (June 2005). "Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling." In: *CoNLL*. Ann Arbor, Michigan, pp. 152–164. URL: https://aclweb.org/anthology/W05-0620.

Caruana, Rich (1997). "Multitask learning." In: *Machine learning* 28.1, pp. 41–75. URL: https://link.springer.com/content/pdf/10.1023/A:1007379606734.pdf.

Cattan, Arie, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan (2020). "Streamlining Cross-Document Coreference Resolution: Evaluation and Modeling." In: *CoRR* abs/2009.11032. arXiv: 2009.11032. URL: https://arxiv.org/abs/2009.11032.

Chang, Kai-Wei, Wen tau Yih, Bishan Yang, and Christopher Meek (2014). "Typed Tensor Decomposition of Knowledge Bases for Relation Extraction." In: *EMNLP*.

Chen, Danqi, Adam Fisch, Jason Weston, and Antoine Bordes (July 2017). "Reading Wikipedia to Answer Open-Domain Questions." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1870–1879. DOI: 10.18653/v1/P17-1171. URL: https://www.aclweb.org/anthology/P17-1171.

Chen, Hong, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong (2018). "PreCo: A Large-scale Dataset in Preschool Vocabulary for Coreference Resolution." In: *EMNLP*. Brussels, Belgium, pp. 172–181. URL: https://aclweb.org/anthology/D18-1016.

Chen, Qian, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang (2016). "Distraction-based neural networks for document summarization." In: *arXiv:1610.08462*.

Cho, Jaemin, Minjoon Seo, and Hannaneh Hajishirzi (Nov. 2019). "Mixture Content Selection for Diverse Sequence Generation." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3121–3131. DOI: 10.

18653/v1/D19-1308. URL: https://www.aclweb.org/anthology/D19-1308.

Choi, Byung-Ju, Jimin Hong, David Park, and Sang Wan Lee (Nov. 2020). "Fˆ2-Softmax: Diversifying Neural Text Generation via Frequency Factorized Softmax." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9167–9182. URL: https://www.aclweb.org/anthology/2020.emnlp-main.737.

Choi, Eunsol, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer (2018). "QuAC: Question Answering in Context." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2174–2184. URL: https://www.aclweb.org/anthology/D18-1241.

Choshen, Leshem, Lior Fox, Zohar Aizenbud, and Omri Abend (2020). "On the Weaknesses of Reinforcement Learning for Neural Machine Translation." In: *ICLR*. URL: https://openreview.net/forum?id=H1eCw3EKvH.

Clark, Jonathan H., Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki (2020). "TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages." In: *Transactions of the Association for Computational Linguistics* 8, pp. 454–470. URL: https://doi.org/10.1162/tacl_a_00317.

Clark, Kevin and Christopher D Manning (2016a). "Deep Reinforcement Learning for Mention-Ranking Coreference Models." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2256–2262.

Clark, Kevin and Christopher D. Manning (Nov. 2016b). "Deep Reinforcement Learning for Mention-Ranking Coreference Models." In: *EMNLP*. Austin, Texas, pp. 2256–2262. URL: https://aclweb.org/anthology/D16-1245.

Clark, Kevin and Christopher D. Manning (Nov. 2016c). "Deep Reinforcement Learning for Mention-Ranking Coreference Models." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2256–2262. DOI: 10.18653/v1/D16-1245. URL: https://www.aclweb.org/anthology/D16-1245.

Clark, Kevin and Christopher D. Manning (Aug. 2016d). "Improving Coreference Resolution by Learning Entity-Level Distributed Representations." In: *ACL*. Berlin, Germany, pp. 643–653. URL: https://aclweb.org/anthology/P16-1061.

Cohen, K. Bretonnel, Karin Verspoor, Karën Fort, Christopher Funk, Michael Bada, Martha Palmer, and Lawrence E. Hunter (2017). "The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain." In: *Handbook of Linguistic Annotation*. Ed. by Nancy Ide and James Pustejovsky. Dordrecht: Springer Netherlands, pp. 1379–1394. ISBN: 978-94-024-

0881-2. DOI: 10.1007/978-94-024-0881-2_53. URL: https://doi.org/10.1007/978-94-024-0881-2_53.

Collins, Liam, Aryan Mokhtari, and Sanjay Shakkottai (2020). "Task-Robust Model-Agnostic Meta-Learning." In: *Advances in Neural Information Processing Systems*. Vol. 33. online. URL: https://arxiv.org/abs/2002.04766.

Cotterell, Ryan and Jason Eisner (July 2017). "Probabilistic Typology: Deep Generative Models of Vowel Inventories." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada, pp. 1182–1192. DOI: 10.18653/v1/P17-1109. URL: https://www.aclweb.org/anthology/P17-1109.

Culotta, Aron, Michael Wick, and Andrew McCallum (2007). "First-Order Probabilistic Models for Coreference Resolution." In: *Proceedings of NAACL*.

Cybulska, Agata and Piek Vossen (May 2014). "Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 4545–4552. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/840_Paper.pdf.

Dai, Bo, Dahua Lin, Raquel Urtasun, and Sanja Fidler (2017). "Towards Diverse and Natural Image Descriptions via a Conditional GAN." In: *CoRR* abs/1703.06029. arXiv: 1703.06029. URL: http://arxiv.org/abs/1703.06029.

Das, Dipanjan and Slav Petrov (2011). "Unsupervised part-of-speech tagging with bilingual graph-based projections." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA, pp. 600–609. URL: https://www.aclweb.org/anthology/P11-1061.pdf.

Day, David, Chad Mchenry, Robyn Kozierok, and Laurel Riek (Jan. 2004). "Callisto: A configurable annotation workbench." In:

Denis, Pascal and Jason Baldridge (2009). "Global joint models for coreference resolution and named entity classification." In: *Procesamiento del Lenguaje Natural* 42.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *NAACL*. Minneapolis, Minnesota, pp. 4171–4186. URL: https://aclweb.org/anthology/N19-1423.

Dhingra, Bhuwan, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen (2019). "Handling Divergent Reference Texts when Evaluating Table-to-Text Generation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 4884–4895.

Dhingra, Bhuwan, Qiao Jin, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov (June 2018). "Neural Models for Reasoning over Multiple Mentions Using Coreference." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 42–48. DOI: 10.18653/v1/N18-2007. URL: https://aclanthology.org/N18-2007.

Dieng, Adji B., Chong Wang, Jianfeng Gao, and John Paisley (2017). "TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency." In: *Proceedings of the 5th International Conference on Learning Representations*. Toulon, France.

Donecker, Paul (1996). "Subdeletion in Verb Phrase Ellipsis." In: *ACL*.

Dong, Li, Jonathan Mallinson, Siva Reddy, and Mirella Lapata (Sept. 2017). "Learning to Paraphrase for Question Answering." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 875–886. DOI: 10.18653/v1/D17-1091. URL: https://www.aclweb.org/anthology/D17-1091.

Dong, Li, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon (2019a). "Unified Language Model Pre-training for Natural Language Understanding and Generation." In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 13042–13054.

Dong, Yue, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung (July 2019b). "EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3393–3402. DOI: 10.18653/v1/P19-1331. URL: https://www.aclweb.org/anthology/P19-1331.

Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://wals.info/.

Du, Xinya, Bhavana Dalvi, Niket Tandon, Antoine Bosselut, Wen-tau Yih, Peter Clark, and Claire Cardie (June 2019). "Be Consistent! Improving Procedural Text Comprehension using Label Consistency." In: *NAACL*. Minneapolis, Minnesota, pp. 2347–2356. URL: https://aclweb.org/anthology/N19-1244.

Durmus, Esin, He He, and Mona Diab (July 2020). "FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5055–5070. DOI: 10.18653/v1/2020.acl-main.454. URL: https://www.aclweb.org/anthology/2020.acl-main.454.

Dziri, Nouha, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane (Aug. 2019). "Augmenting Neural Response Generation with

Context-Aware Topical Attention." In: *Proceedings of the First Workshop on NLP for Conversational AI*. Florence, Italy: Association for Computational Linguistics, pp. 18–31. DOI: 10.18653/v1/W19-4103. URL: https://www.aclweb.org/anthology/W19-4103.

Elango, Pradheep (2005). "Coreference resolution: A survey." In: *University of Wisconsin, Madison, WI*.

Elbourne, Paul (2013). *Definite descriptions*. Vol. 1. Oxford Studies in Semantics.

Emami, Ali, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung (June 2018a). "A Generalized Knowledge Hunting Framework for the Winograd Schema Challenge." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. New Orleans, Louisiana, USA: Association for Computational Linguistics, pp. 25–31. DOI: 10.18653/v1/N18-4004. URL: https://www.aclweb.org/anthology/N18-4004.

Emami, Ali, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung (June 2018b). "A Generalized Knowledge Hunting Framework for the Winograd Schema Challenge." In: *NAACL*. New Orleans, Louisiana, USA, pp. 25–31. URL: https://aclweb.org/anthology/N18-4004.

Falke, Tobias, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych (2019). "Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 2214–2220.

Fan, Angela, Mike Lewis, and Yann Dauphin (July 2018). "Hierarchical Neural Story Generation." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 889–898. DOI: 10.18653/v1/P18-1082. URL: https://www.aclweb.org/anthology/P18-1082.

Fellbaum, Christiane, ed. (1998). *WordNet: an electronic lexical database*. MIT Press.

Finkel, Jennifer and Chris Manning (2008). "Enforcing Transitivity in Coreference Resolution." In: *Proceedings of ACL*.

Finn, Chelsea, Pieter Abbeel, and Sergey Levine (2017). "Model-agnostic meta-learning for fast adaptation of deep networks." In: *Proceedings of the 34th International Conference on Machine Learning*. Sydney, Australia, pp. 1126–1135. URL: http://proceedings.mlr.press/v70/finn17a/finn17a.pdf.

Finn, Chelsea, Kelvin Xu, and Sergey Levine (2018). "Probabilistic Model-Agnostic Meta-Learning." In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Montreal, Canada, pp. 9516–9527. URL: https://proceedings.neurips.cc/paper/2018/file/8e2c381d4dd04f1c55093f22c59c3a08-Paper.pdf.

Freitag, Markus, David Grangier, and Isaac Caswell (Nov. 2020). "BLEU might be Guilty but References are not Innocent." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 61–71. URL: https://www.aclweb.org/anthology/2020.emnlp-main.5.

Gabriel, Saadia, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao (2020). *Go Figure! A Meta Evaluation of Factuality in Summarization*. arXiv: 2010.12834 [cs.CL].

Gardner, Matt, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min (2019). *Question Answering is a Format; When is it Useful?* arXiv: 1909.11291 [cs.CL].

Ge, Niyu, John Hale, and Eugene Charniak (1998). "A Statistical Approach to Anaphora Resolution." In: *Sixth Workshop on Very Large Corpora*. URL: https://aclanthology.org/W98-1119.

Gehrmann, Sebastian, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur P. Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou (2021). "The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics." In: *CoRR* abs/2102.01672. URL: https://arxiv.org/abs/2102.01672.

Gehrmann, Sebastian, Yuntian Deng, and Alexander Rush (2018). "Bottom-Up Abstractive Summarization." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4098–4109.

Gemp, Ian and Sridhar Mahadevan (2018). "Global convergence to the equilibrium of GANs using variational inequalities." In: *arXiv preprint arXiv:1808.01531*. URL: https://arxiv.org/pdf/1808.01531.pdf.

Gerz, Daniela, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen (2018a). "On the Relation between Linguistic Typology and (Limitations of) Multilingual Language Modeling." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 316–327. DOI: 10.18653/v1/D18-1029. URL: https://www.aclweb.org/anthology/D18-1029.

Gerz, Daniela, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen (2018b). "Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction." In: *Transactions of the Association for Computational Linguistics* 6, pp. 451–465. URL: https://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00032.

Geva, Mor, Eric Malmi, Idan Szpektor, and Jonathan Berant (2019). "DiscoFuse: A Large-Scale Dataset for Discourse-Based Sentence Fusion." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3443–3455.

Ghaddar, Abbas and Philippe Langlais (2016). "WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA). Portorož, Slovenia: European Language Resources Association (ELRA). ISBN: 978-2-9517408-9-1.

Ghosh, Shalini, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck (2016). "Contextual LSTM (CLSTM) models for Large scale NLP tasks." In: *CoRR* abs/1602.06291.

Globerson, Amir and Sam Roweis (2006). "Nightmare at Test Time: Robust Learning by Feature Deletion." In: *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, Pennsylvania, USA, 353–360. ISBN: 1595933832. DOI: 10.1145/1143844.1143889. URL: https://doi.org/10.1145/1143844.1143889.

Grant, Erin, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths (2018). "Recasting Gradient-Based Meta-Learning as Hierarchical Bayes." In: *International Conference on Learning Representations*. Vancouver, Canada. URL: https://openreview.net/forum?id=BJ_UL-k0b.

Graves, Alex and Jürgen Schmidhuber (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." In: *Neural Networks* 18.5-6, pp. 602–610.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein (1995). "Centering: A Framework for Modeling the Local Coherence of Discourse." In: *Computational Linguistics* 21.2, pp. 203–225. URL: https://aclanthology.org/J95-2003.

Gu, Jiatao, Zhengdong Lu, Hang Li, and Victor O.K. Li (Aug. 2016). "Incorporating Copying Mechanism in Sequence-to-Sequence Learning." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1631–1640. DOI: 10.18653/v1/P16-1154. URL: https://www.aclweb.org/anthology/P16-1154.

Gu, Jiatao, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho (2018). "Meta-Learning for Low-Resource Neural Machine Translation." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, pp. 3622–

3631. DOI: 10.18653/v1/D18-1398. URL: https://www.aclweb.org/anthology/D18-1398.

Hacioglu, Kadri (2004). "Semantic role labeling using dependency trees." In: *COLING*, pp. 1273–1276.

Haghighi, Aria and Dan Klein (2009). "Simple coreference resolution with rich syntactic and semantic features." In: *EMNLP*. ACL, pp. 1152–1161.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank, eds. (2016). *Glottolog 2.7*. Jena: Max Planck Institute for the Science of Human History. URL: http://glottolog.org.

Hashimoto, Tatsunori, Hugh Zhang, and Percy Liang (June 2019). "Unifying Human and Statistical Evaluation for Natural Language Generation." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1689–1701. DOI: 10.18653/v1/N19-1169. URL: https://www.aclweb.org/anthology/N19-1169.

He, Luheng, Kenton Lee, Omer Levy, and Luke Zettlemoyer (July 2018). "Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling." In: *ACL*. Melbourne, Australia, pp. 364–369. URL: https://aclweb.org/anthology/P18-2058.

He, Luheng, Kenton Lee, Mike Lewis, and Luke Zettlemoyer (July 2017). "Deep Semantic Role Labeling: What Works and What's Next." In: *ACL*. Vancouver, Canada, pp. 473–483. URL: https://aclweb.org/anthology/P17-1044.

He, Luheng, Mike Lewis, and Luke Zettlemoyer (2015). "Question-answer driven semantic role labeling: Using natural language to annotate natural language." In: *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 643–653.

Hermann, Karl Moritz, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom (2015). "Teaching Machines to Read and Comprehend." In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., pp. 1693–1701.

Hershcovich, Daniel, Omri Abend, and Ari Rappoport (July 2017). "A Transition-Based Directed Acyclic Graph Parser for UCCA." In: *ACL*. Vancouver, Canada, pp. 1127–1138. URL: https://aclweb.org/anthology/P17-1104.

Hobbs, J (1986). "Resolving Pronoun References." In: *Readings in Natural Language Processing*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 339–352. ISBN: 0934613117.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997a). "Long Short-Term Memory." In: *Neural Comput.* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hochreiter, Sepp and Jürgen Schmidhuber (1997b). "Long short-term memory." In: *Neural computation* 9.8, pp. 1735–1780.

Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi (2020). "The Curious Case of Neural Text Degeneration." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rygGQyrFvH.

Hou, Yufang (2020). *Bridging Anaphora Resolution as Question Answering*. arXiv: 2004.07898 [cs.CL].

Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson (2020). "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation." In: *Proceedings of the 37th International Conference on Machine Learning*. online, pp. 4411–4421. URL: http://proceedings.mlr.press/v119/hu20b/hu20b.pdf.

Jin, Chi, Praneeth Netrapalli, and Michael Jordan (2020). "What is local optimality in nonconvex–nonconcave minimax optimization?" In: *Proceedings of the 37th International Conference on Machine Learning*. online, pp. 4880–4889. URL: http://proceedings.mlr.press/v119/jin20e/jin20e.pdf.

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy (2019a). *SpanBERT: Improving Pre-training by Representing and Predicting Spans*. arXiv: 1907.10529 [cs.CL].

Joshi, Mandar, Omer Levy, Luke Zettlemoyer, and Daniel Weld (2019b). "BERT for Coreference Resolution: Baselines and Analysis." In: *EMNLP*.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (July 2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. online, pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560. URL: https://www.aclweb.org/anthology/2020.acl-main.560.

Junczys-Dowmunt, Marcin (2019). "Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation." In: *arXiv:1907.06170*.

Kamp, Hans and Uwe Reyle (2013). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Springer.

Karmaker Santu, Shubhra Kanti, Kalyan Veeramachaneni, and Chengxiang Zhai (Nov. 2019). "TILM: Neural Language Models with Evolving Topical Influence." In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 778–788. DOI: 10.18653/v1/K19-1073. URL: https://www.aclweb.org/anthology/K19-1073.

Khashabi, Daniel, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth (June 2018). "Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 252–

262. DOI: 10.18653/v1/N18-1023. URL: https://aclanthology.org/N18-1023.

Kim, Byeongchang, Hyunwoo Kim, and Gunhee Kim (2019). "Abstractive Summarization of Reddit Posts with Multi-level Memory Networks." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2519–2531.

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization." In: *CoRR* abs/1412.6980. arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization." In: *International Conference on Learning Representations*. Ed. by Yoshua Bengio and Yann LeCun. San Diego, California, USA. URL: http://arxiv.org/abs/1412.6980.

Kipf, Thomas N. and Max Welling (2017). "Semi-Supervised Classification with Graph Convolutional Networks." In: *ICLR*. OpenReview.net. URL: https://openreview.net/forum?id=SJU4ayYgl.

Kong, Fang, GuoDong Zhou, and Qiaoming Zhu (Aug. 2009). "Employing the Centering Theory in Pronoun Resolution from the Semantic Perspective." In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 987–996. URL: https://aclanthology.org/D09-1103.

Kryscinski, Wojciech, Bryan McCann, Caiming Xiong, and Richard Socher (2019). "Evaluating the Factual Consistency of Abstractive Text Summarization." In: *CoRR* abs/1910.12840.

Kryscinski, Wojciech, Bryan McCann, Caiming Xiong, and Richard Socher (Nov. 2020). "Evaluating the Factual Consistency of Abstractive Text Summarization." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9332–9346. URL: https://www.aclweb.org/anthology/2020.emnlp-main.750.

Kryściński, Wojciech, Bryan McCann, Caiming Xiong, and Richard Socher (2019). *Evaluating the Factual Consistency of Abstractive Text Summarization*. arXiv: 1910.12840 [cs.CL].

Kudo, Taku and John Richardson (2018). "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 66–71.

Kulikov, Ilia, Alexander Miller, Kyunghyun Cho, and Jason Weston (2019). "Importance of Search and Evaluation Strategies in Neural Dialogue Modeling." In: *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo, Japan: Association for Computational Linguistics, pp. 76–87. DOI: 10.18653/v1/W19-8609. URL: https://www.aclweb.org/anthology/W19-8609.

Lai, Guokun, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy (Sept. 2017). "RACE: Large-scale ReAding Comprehension Dataset From Examinations." In: *EMNLP*. Copenhagen, Denmark, pp. 785–794. URL: https://aclweb.org/anthology/D17-1082.

Langford, John and Bianca Zadrozny (2005). "Relating Reinforcement Learning Performance to Classification Performance." In: *ICML*.

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (2017a). "End-to-end Neural Coreference Resolution." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 188–197.

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (Sept. 2017b). "End-to-end Neural Coreference Resolution." In: *EMNLP*. Copenhagen, Denmark, pp. 188–197. URL: https://aclweb.org/anthology/D17-1018.

Lee, Kenton, Luheng He, and Luke Zettlemoyer (2018a). "Higher-Order Coreference Resolution with Coarse-to-Fine Inference." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 687–692.

Lee, Kenton, Luheng He, and Luke Zettlemoyer (June 2018b). "Higher-Order Coreference Resolution with Coarse-to-Fine Inference." In: *NAACL*. New Orleans, Louisiana, pp. 687–692. URL: https://aclweb.org/anthology/N18-2108.

Lee, Kenton, Luheng He, and Luke Zettlemoyer (June 2018c). "Higher-Order Coreference Resolution with Coarse-to-Fine Inference." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 687–692. DOI: 10.18653/v1/N18-2108. URL: https://www.aclweb.org/anthology/N18-2108.

Lerer, Adam, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich (2019). "PyTorch-BigGraph: A Large-scale Graph Embedding System." In: *Proceedings of the 2nd SysML Conference*. Palo Alto, CA, USA.

Letcher, Alistair, David Balduzzi, Sébastien Racaniere, James Martens, Jakob N Foerster, Karl Tuyls, and Thore Graepel (2019). "Differentiable Game Mechanics." In: *Journal of Machine Learning Research* 20, pp. 84–1. URL: https://www.jmlr.org/papers/volume20/19-008/19-008.pdf.

Levin, Beth (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Levy, Omer, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer (2017). "Zero-Shot Relation Extraction via Reading Comprehension." In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 333–342. DOI: 10.18653/v1/K17-1034. URL: http://aclweb.org/anthology/K17-1034.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettle-

moyer (2019). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: *CoRR* abs/1910.13461.

Li, Jiwei, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan (June 2016a). "A Diversity-Promoting Objective Function for Neural Conversation Models." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 110–119. DOI: 10.18653/v1/N16-1014. URL: https://www.aclweb.org/anthology/N16-1014.

Li, Jiwei, Will Monroe, and Dan Jurafsky (2016b). "A Simple, Fast Diverse Decoding Algorithm for Neural Generation." In: *CoRR* abs/1611.08562. arXiv: 1611.08562. URL: http://arxiv.org/abs/1611.08562.

Li, Xiaoya, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li (July 2020a). "A Unified MRC Framework for Named Entity Recognition." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5849–5859. DOI: 10.18653/v1/2020.acl-main.519. URL: https://www.aclweb.org/anthology/2020.acl-main.519.

Li, Zheng, Mukul Kumar, William Headden, Bing Yin, Ying Wei, Yu Zhang, and Qiang Yang (Nov. 2020b). "Learn to Cross-lingual Transfer with Meta Graph Learning Across Heterogeneous Languages." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. online, pp. 2290–2301. URL: https://www.aclweb.org/anthology/2020.emnlp-main.179.

Li, Zhenguo, Fengwei Zhou, Fei Chen, and Hang Li (2017). "Meta-SGD: Learning to learn quickly for few-shot learning." In: *arXiv preprint arXiv:1707.09835*. URL: https://arxiv.org/pdf/1707.09835.pdf.

Lin, Chin-Yew and Eduard Hovy (2000). "The Automated Acquisition of Topic Signatures for Text Summarization." In: *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*. URL: https://www.aclweb.org/anthology/C00-1072.

Lin, Chin Yew and Eduard Hovy (2003). "Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics." In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 150–157.

Lin, Chuan-Jie, Chien-Wei Pao, Yen-Heng Chen, Chi-Ting Liu, and Hui-Huang Hsu (2016). "Ellipsis and Coreference Resolution in a Computerized Virtual Patient Dialogue System." In: *Journal of Medical Systems* 40, p. 206.

Lin, Yankai, Zhiyuan Liu, and Maosong Sun (2015). "Modeling Relation Paths for Representation Learning of Knowledge Bases." In: *EMNLP*.

Liu, Jiangming, Shay B. Cohen, and Mirella Lapata (July 2018). "Discourse Representation Structure Parsing." In: *ACL*. Melbourne, Australia, pp. 429–439. URL: https://aclweb.org/anthology/P18-1040.

Liu, Yang and Mirella Lapata (2018). "Learning Structured Text Representations." In: *TACL*, pp. 63–75. URL: https://aclweb.org/anthology/Q18-1005.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." In: *CoRR* abs/1907.11692.

Mallinson, Jonathan, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido (2020). *Felix: Flexible Text Editing Through Tagging and Insertion*. arXiv: 2003.10687 [cs.CL].

Malmi, Eric, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn (Nov. 2019). "Encode, Tag, Realize: High-Precision Text Editing." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 5054–5065. DOI: 10.18653/v1/D19-1510. URL: https://www.aclweb.org/anthology/D19-1510.

Mani, Inderjeet (2001). *Automatic summarization*. Vol. 3. John Benjamins Publishing.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit." In: *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60. URL: http://www.aclweb.org/anthology/P/P14/P14-5010.

Mao, Yuning, Xiang Ren, Jiaming Shen, Xiaotao Gu, and Jiawei Han (2018). "Building a large-scale annotated Chinese corpus." In: *ACL*.

Markert, Katja, Malvina Nissim, and Natalia Modjeska (2003). "Using the Web for Nominal Anaphora Resolution." In: *Proceedings of the 2003 EACL Workshop on The Computational Treatment of Anaphora*. URL: https://aclanthology.org/W03-2606.

Martins, Pedro Henrique, Zita Marinho, and André F. T. Martins (Nov. 2020). "Sparse Text Generation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4252–4273. URL: https://www.aclweb.org/anthology/2020.emnlp-main.348.

Martschat, Sebastian and Michael Strube (Oct. 2014). "Recall Error Analysis for Coreference Resolution." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 2070–2081. DOI: 10.3115/v1/D14-1221. URL: https://www.aclweb.org/anthology/D14-1221.

Maruf, Sameen and Gholamreza Haffari (July 2018). "Document Context Neural Machine Translation with Memory Networks." In:

ACL. Melbourne, Australia, pp. 1275–1284. URL: https://aclweb.org/anthology/P18-1118.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald (July 2020). "On Faithfulness and Factuality in Abstractive Summarization." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: https://www.aclweb.org/anthology/2020.acl-main.173.

McCallum, Andrew and Ben Wellner (2005). "Conditional Models of Identity Uncertainty with Application to Noun Coreference." In: *Advances in Neural Information Processing Systems*. Ed. by L. Saul, Y. Weiss, and L. Bottou. Vol. 17. MIT Press. URL: https://proceedings.neurips.cc/paper/2004/file/1680829293f2a8541-efa2647a0290f88-Paper.pdf.

McCann, Bryan, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher (2018a). *The Natural Language Decathlon: Multitask Learning as Question Answering*. arXiv: 1806.08730 [cs.CL].

McCann, Bryan, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher (2018b). "The natural language decathlon: Multitask learning as question answering." In: *arXiv preprint arXiv:1806.08730*.

Meng, Yuanliang and Anna Rumshisky (2018). "Triad-based Neural Network for Coreference Resolution." In: *COLING*.

Merchant, Jason (2001). *The syntax of silence: Sluicing, islands, and the theory of ellipsis*. Oxford University Press on Demand.

Mikolov, Tomas and Geoffrey Zweig (2012). "Context dependent recurrent neural network language model." In: *Proceedings of the Spoken Language Technology Workshop*. IEEE, pp. 234–239.

Miller, George A. (Nov. 1995). "WordNet: A Lexical Database for English." In: *Commun. ACM* 38.11, 39–41. ISSN: 0001-0782. DOI: 10.1145/219717.219748. URL: https://doi.org/10.1145/219717.219748.

Müller, Christoph and Michael Strube (2006). "Multi-level annotation of linguistic data with MMAX2." In: *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Ed. by Sabine Braun, Kurt Kohn, and Joybrato Mukherjee. Frankfurt a.M., Germany: Peter Lang, pp. 197–214.

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (2018). "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1797–1807.

Narayan, Shashi, Siva Reddy, and Shay B. Cohen (2016). "Paraphrase Generation from Latent-Variable PCFGs for Semantic Parsing." In: *Proceedings of the 9th International Natural Language Generation conference*. Edinburgh, UK: Association for Computational Linguistics, pp. 153–162. DOI: 10.18653/v1/W16-6625. URL: https://www.aclweb.org/anthology/W16-6625.

Nenkova, Ani and Kathleen McKeown (2011). "Automatic Summarization." In: *Foundations and Trends in Information Retrieval* 5.2–3, pp. 103–233.

Ng, Vincent (2007). "Shallow Semantics for Coreference Resolution." In: *IJCAI*. Vol. 2007, pp. 1689–1694.

Nghiem, Minh-Quoc and Sophia Ananiadou (2018). "APLenty: annotation tool for creating high-quality datasets using active and proactive learning." In: *Proceedings of EMNLP*.

Nooralahzadeh, Farhad, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein (Nov. 2020). "Zero-Shot Cross-Lingual Transfer with Meta Learning." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. online, pp. 4547–4562. URL: https://www.aclweb.org/anthology/2020.emnlp-main.368.

O'Gorman, Tim, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer (Aug. 2018). "AMR Beyond the Sentence: the Multi-sentence AMR corpus." In: *COLING*. Santa Fe, New Mexico, USA, pp. 3693–3702. URL: https://aclweb.org/anthology/C18-1313.

Orbanz, Peter (2012). "Lecture notes on Bayesian nonparametrics." In: *Journal of Mathematical Psychology* 56, pp. 1–12. URL: http://www.gatsby.ucl.ac.uk/~porbanz/papers/porbanz_BNP_draft.pdf.

Oren, Yonatan, Shiori Sagawa, Tatsunori Hashimoto, and Percy Liang (Nov. 2019). "Distributionally Robust Language Modeling." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 4227–4237. DOI: 10.18653/v1/D19-1432. URL: https://www.aclweb.org/anthology/D19-1432.

Partee, Barbara (1978). "Bound variables and other anaphors." In: *Theoretical Issues In Natural Language Processing*. Blackwell.

Pasunuru, Ramakanth and Mohit Bansal (2018). "Multi-Reward Reinforced Summarization with Saliency and Entailment." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 646–653.

Pearlmutter, Barak A. (1994). "Fast exact multiplication by the Hessian." In: *Neural computation* 6.1, pp. 147–160. URL: https://www.mitpressjournals.org/doi/pdfplus/10.1162/neco.1994.6.1.147.

Peng, Haoruo, Daniel Khashabi, and Dan Roth (2015). "Solving Hard Coreference Problems." In: *Proceedings of NAACL*.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014a). "GloVe: Global Vectors for Word Representation." In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014b). "Glove: Global Vectors for Word Representation." In: *EMNLP*.

Doha, Qatar, pp. 1532–1543. URL: https://aclweb.org/anthology/D14-1162.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep Contextualized Word Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237.

Poesio, Massimo, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi (Apr. 2013a). "Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation." In: *ACM Trans. Interact. Intell. Syst.* 1.

Poesio, Massimo, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi (Apr. 2013b). "Phrase Detectives: Utilizing Collective Intelligence for Internet-scale Language Resource Creation." In: *ACM Trans. Interact. Intell. Syst.* 3.1, 3:1–3:44. ISSN: 2160-6455. DOI: 10.1145/2448116.2448119. URL: http://doi.acm.org/10.1145/2448116.2448119.

Poesio, Massimo, Rahul Mehta, Axel Maroudas, and Janet Hitzeman (2004). "Learning to Resolve Bridging References." In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. ACL '04. Barcelona, Spain: Association for Computational Linguistics, 143–es. DOI: 10.3115/1218955.1218974. URL: https://doi.org/10.3115/1218955.1218974.

Ponti, Edoardo M., Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen (2021). "Parameter Space Factorization for Zero-Shot Learning across Tasks and Languages." In: *Transactions of the Association for Computational Linguistics* 9, pp. 410–428. URL: http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00374/1912912/tacl_a_00374.pdf.

Ponti, Edoardo Maria, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen (Nov. 2020). "XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. online, pp. 2362–2376. DOI: 10.18653/v1/2020.emnlp-main.185. URL: https://www.aclweb.org/anthology/2020.emnlp-main.185.

Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen (2019a). "Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing." In: *Computational Linguistics* 45.3, pp. 559–601. URL: https://arxiv.org/pdf/1807.00914.pdf.

Ponti, Edoardo Maria, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen (Nov. 2019b). "Towards Zero-shot Language Modeling." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 2900–2910. DOI: 10.18653/v1/D19-1288. URL: https://www.aclweb.org/anthology/D19-1288.

Ponti, Edoardo (2021). "Inductive Bias and Modular Design for Sample-Efficient Neural Language Learning." PhD thesis. University of Cambridge. URL: https://aspace.repository.cam.ac.uk/bitstream/handle/1810/319303/thesis_electronic.pdf.

Ponzetto, Simone Paolo and Michael Strube (June 2006a). "Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution." In: *NAACL*. New York City, USA, pp. 192–199. URL: https://aclweb.org/anthology/N06-1025.

Ponzetto, Simone Paolo and Michael Strube (2006b). "Semantic Role Labeling for Coreference Resolution." In: *Demonstrations*. URL: https://aclweb.org/anthology/E06-2015.

Postal, Paul (1966). "On so-called pronouns in English." In: *Monograph series on language and linguistics* 19, pp. 177–206.

Pradhan, Sameer, Kadri Hacioglu, Wayne Ward, James H Martin, and Dan Jurafsky (2005). "Semantic role chunking combining complementary syntactic views." In: *CoNLL*, pp. 217–220.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang (2012a). "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes." In: *Joint Conference on EMNLP and CoNLL - Shared Task*. Association for Computational Linguistics, pp. 1–40. URL: http://www.aclweb.org/anthology/W12-4501.

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang (July 2012b). "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes." In: *CoNLL*. Jeju Island, Korea, pp. 1–40. URL: https://aclweb.org/anthology/W12-4501.

Prange, Jakob, Nathan Schneider, and Omri Abend (Aug. 2019). "Semantically Constrained Multilayer Annotation: The Case of Coreference." In: *DMR*. Florence, Italy, pp. 164–176. URL: https://aclweb.org/anthology/W19-3319.

Pratt, Lorien Y (1993). "Discriminability-based transfer between neural networks." In: *Advances in neural information processing systems*. Vol. 5, pp. 204–204. URL: https://proceedings.neurips.cc/paper/1992/file/67e103b0761e60683e83c559be18d40c-Paper.pdf.

Prokofyev, Roman, Alberto Tonon, Michael Luggen, Loic Vouilloz, Djellel Eddine Difallah, and Philippe Cudré-Mauroux (2015). "SANAPHOR: Ontology-Based Coreference Resolution." In: *International Semantic Web Conference*.

Punyakanok, Vasin, Dan Roth, and Wen-tau Yih (2008). "The importance of syntactic parsing and inference in semantic role labeling." In: *Computational Linguistics* 2, pp. 257–287.

Qi, Weizhen, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou (Nov. 2020). "ProphetNet: Predicting Future N-gram for Sequence-to-SequencePre-training." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2401–2410. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.217.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). "Improving language understanding by generative pre-training." In: *Technical report, OpenAI*.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2019a). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. arXiv: 1910.10683 [cs.LG].

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2019b). "Exploring the limits of transfer learning with a unified text-to-text transformer." In: *arXiv preprint arXiv:1910.10683*.

Rahman, Altaf and Vincent Ng (June 2011a). "Coreference Resolution with World Knowledge." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 814–824. URL: https://aclanthology.org/P11-1082.

Rahman, Altaf and Vincent Ng (2011b). "Coreference resolution with world knowledge." In: *ACL*. ACL, pp. 814–824.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (Nov. 2016a). "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. URL: https://www.aclweb.org/anthology/D16-1264.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016b). "Squad: 100,000+ questions for machine comprehension of text." In: *arXiv preprint arXiv:1606.05250*.

Reddy, Siva, Danqi Chen, and Christopher D. Manning (Mar. 2019). "CoQA: A Conversational Question Answering Challenge." In: *Transactions of the Association for Computational Linguistics* 7, pp. 249–266. DOI: 10.1162/tacl_a_00266. URL: https://aclanthology.org/Q19-1016.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie (Nov. 2020). "COMET: A Neural Framework for MT Evaluation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213. URL: https://www.aclweb.org/anthology/2020.emnlp-main.213.

Riedel, Sebastian, Limin Yao, Andrew McCallum, and Benjamin M. Marlin (June 2013). "Relation Extraction with Matrix Factorization and Universal Schemas." In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 74–84. URL: https://www.aclweb.org/anthology/N13-1008.

Rønning, Ola, Daniel Hardt, and Anders Søgaard (2018). "Sluice Resolution without Hand-Crafted Features over Brittle Syntax Trees." In: *Proceedings of the 2018 Conference of the North American Chap-*

ter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, 236–241. URL: http://www.aclweb.org/anthology/N18-2038.

Rothe, Sascha, Shashi Narayan, and Aliaksei Severyn (2020). "Leveraging pre-trained checkpoints for sequence generation tasks." In: *Transactions of the Association for Computational Linguistics* 8, pp. 264–280.

Ruder, Sebastian (2019). "Neural transfer learning for natural language processing." PhD thesis. NUI Galway. URL: https://aran.library.nuigalway.ie/bitstream/handle/10379/15463/neural_transfer_learning_for_nlp.pdf.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard (2019). "A survey of cross-lingual embedding models." In: *Journal of Artificial Intelligence Research* 65, pp. 569–631. URL: https://doi.org/10.1613/jair.1.11640.

Runner, Jeffrey T, Rachel S Sussman, and Michael K Tanenhaus (2003). "Assignment of reference to reflexives and pronouns in picture noun phrases: evidence from eye movements." In: *Cognition* 89.1, B1 –B13. ISSN: 0010-0277. DOI: https://doi.org/10.1016/S0010-0277(03)00065-9. URL: http://www.sciencedirect.com/science/article/pii/S0010027703000659.

Sadek, Jawad and Farid Meziane (2016). "A discourse-based approach for Arabic question answering." In: *TALLIP* 2, pp. 1–18.

Schäfer, Florian and Anima Anandkumar (2019). "Competitive gradient descent." In: *Advances in Neural Information Processing Systems*. Vol. 32. Vancouver, Canada, pp. 7625–7635. URL: https://proceedings.neurips.cc/paper/2019/file/56c51a39a7c77d80-84838cc920585bd0-Paper.pdf.

Schoch, Stephanie, Diyi Yang, and Yangfeng Ji (Dec. 2020). ""This is a Problem, Don't You Agree?" Framing and Bias in Human Evaluation for Natural Language Generation." In: *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*. Online (Dublin, Ireland): Association for Computational Linguistics, pp. 10–16. URL: https://www.aclweb.org/anthology/2020.evalnlgeval-1.2.

See, Abigail, Peter J. Liu, and Christopher D. Manning (2017). "Get To The Point: Summarization with Pointer-Generator Networks." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada: Association for Computational Linguistics, pp. 1073–1083.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh (July 2020). "BLEURT: Learning Robust Metrics for Text Generation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7881–7892. DOI: 10.18653/v1/2020.acl-main.704. URL: https://www.aclweb.org/anthology/2020.acl-main.704.

Shibata, Tomohide and Sadao Kurohashi (July 2018). "Entity-Centric Joint Modeling of Japanese Coreference Resolution and Predicate Argument Structure Analysis." In: *ACL*. Melbourne, Australia, pp. 579–589. URL: https://aclweb.org/anthology/P18-1054.

Silveira, Natalia, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning (May 2014). "A Gold Standard Dependency Corpus for English." In: *LREC*. Reykjavik, Iceland, pp. 2897–2904. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1089_Paper.pdf.

Song, Kaiqiang, Logan Lebanoff, Q. Guo, Xipeng Qiu, X. Xue, Chen Li, Dong Yu, and Fei Liu (2020). "Joint Parsing and Generation for Abstractive Summarization." In: *AAAI*.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu (2019). "MASS: Masked Sequence to Sequence Pre-training for Language Generation." In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. PMLR, pp. 5926–5936.

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim (2001). "A Machine Learning Approach to Coreference Resolution of Noun Phrases." In: *Computational Linguistics* 27.4, pp. 521–544. DOI: 10.1162/089120101753342653. URL: https://aclanthology.org/J01-4004.

Staggers, Nancy and A. F. Norcio (Apr. 1993). "Mental Models: Concepts for Human-Computer Interaction Research." In: *Int. J. Man-Mach. Stud.* 38.4, 587–605. ISSN: 0020-7373. DOI: 10.1006/imms.1993.1028. URL: https://doi.org/10.1006/imms.1993.1028.

Strube, Michael (Aug. 1998). "Never Look Back: An Alternative to Centering." In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*. Montreal, Quebec, Canada: Association for Computational Linguistics, pp. 1251–1257. DOI: 10.3115/980691.980773. URL: https://aclanthology.org/P98-2204.

Strubell, Emma, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum (2018). "Linguistically-Informed Self-Attention for Semantic Role Labeling." In: *EMNLP*. Brussels, Belgium, pp. 5027–5038. URL: https://www.aclweb.org/anthology/D18-1548.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum (2007). "Yago: A Core of Semantic Knowledge." In: *Proceedings of the 16th International Conference on World Wide Web*. WWW '07. Banff, Alberta, Canada: Association for Computing Machinery, 697–706. ISBN: 9781595936547. DOI: 10.1145/1242572.1242667. URL: https://doi.org/10.1145/1242572.1242667.

Sultan, Md Arafat, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli (July 2020). "On the Importance of Diversity in Question Generation for QA." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5651–5656. DOI: 10.18653/v1/2020.acl-main.500. URL: https://www.aclweb.org/anthology/2020.acl-main.500.

Sutton, Charles and Andrew McCallum (2005). *Joint parsing and semantic role labeling*. Tech. rep. MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE.

Tang, Duyu, Bing Qin, and Ting Liu (Aug. 2015). "Learning Semantic Representations of Users and Products for Document Level Senti-

ment Classification." In: *ACL*. Beijing, China, pp. 1014–1023. URL: https://aclweb.org/anthology/P15-1098/.

Teh, Yee, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu (2017). "Distral: Robust multitask reinforcement learning." In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 4496–4506. URL: http://papers.nips.cc/paper/7036-distral-robust-multitask-reinforcement-learning.pdf.

Tetreault, Joel R. (June 1999). "Analysis of Syntax-Based Pronoun Resolution Methods." In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland, USA: Association for Computational Linguistics, pp. 602–605. DOI: 10.3115/1034678.1034688. URL: https://aclanthology.org/P99-1079.

Tian, Ran, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh (2019). *Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation*. arXiv: 1910.08684 [cs.CL].

Tiedemann, Jörg (Aug. 2015). "Cross-Lingual Dependency Parsing with Universal Dependencies and Predicted PoS Labels." In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala, Sweden, pp. 340–349. URL: https://www.aclweb.org/anthology/W15-2137.

Toutanova, Kristina, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon (2015). "Representing Text for Joint Embedding of Text and Knowledge Bases." In: *EMNLP*.

Turc, Iulia, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). *Well-Read Students Learn Better: On the Importance of Pre-training Compact Models*. arXiv: 1908.08962 [cs.CL].

Uryupina, Olga (May 2006). "Coreference Resolution with and without Linguistic Knowledge." In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/726_pdf.pdf.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need." In: *Advances in Neural Information Processing Systems 30*, pp. 5998–6008.

Veličković, Petar, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm (2018). "Deep graph infomax." In: *arXiv:1809.10341*.

Verberne, Suzan, LWJ Boves, NHJ Oostdijk, and PAJM Coppen (2007). "Discourse-based answering of why-questions." In: *TAL*.

Verga, Patrick, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum (June 2016). "Multilingual Relation Extraction using Compositional Universal Schema." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego,

California: Association for Computational Linguistics, pp. 886–896. DOI: 10.18653/v1/N16-1103. URL: https://www.aclweb.org/anthology/N16-1103.

Verga, Patrick and Andrew McCallum (2016a). "Row-less Universal Schema." In: *AKBC*.

Verga, Patrick and Andrew McCallum (2016b). "Row-less universal schema." In: *arXiv preprint arXiv:1604.06361*.

Vijayakumar, Ashwin K., Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra (2018). "Diverse Beam Search for Improved Description of Complex Scenes." In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. AAAI Press, pp. 7371–7379. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17329.

Voita, Elena, Pavel Serdyukov, Rico Sennrich, and Ivan Titov (July 2018). "Context-Aware Neural Machine Translation Learns Anaphora Resolution." In: *ACL*. Melbourne, Australia, pp. 1264–1274. URL: https://aclweb.org/anthology/P18-1117.

Wagstaff, Kiri Lou and Claire Cardie (2002). "Intelligent Clustering with Instance-Level Constraints." AAI3059148. PhD thesis. USA. ISBN: 0493751823.

Wang, Alex, Kyunghyun Cho, and Mike Lewis (July 2020a). "Asking and Answering Questions to Evaluate the Factual Consistency of Summaries." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5008–5020. DOI: 10.18653/v1/2020.acl-main.450. URL: https://www.aclweb.org/anthology/2020.acl-main.450.

Wang, Sinong, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma (2020b). "Linformer: Self-Attention with Linear Complexity." In: *CoRR* abs/2006.04768. arXiv: 2006.04768. URL: https://arxiv.org/abs/2006.04768.

Wang, Xinyi, Yulia Tsvetkov, and Graham Neubig (July 2020c). "Balancing Training for Multilingual Neural Machine Translation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. online, pp. 8526–8537. DOI: 10.18653/v1/2020.acl-main.754. URL: https://www.aclweb.org/anthology/2020.acl-main.754.

Wang, Zhen, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang (Nov. 2020d). "Diversify Question Generation with Continuous Content Selectors and Question Type Modeling." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2134–2143. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.194.

Wang, Zhengjue, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou (Nov. 2020e). "Friendly Topic As-

sistant for Transformer Based Abstractive Summarization." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 485–497. URL: https://www.aclweb.org/anthology/2020.emnlp-main.35.

Wang, Zhenyi, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen (2020f). *Towards Faithful Neural Table-to-Text Generation with Content-Matching Constraints*. arXiv: 2005.00969 [cs.CL].

Webster, Kellie, Marta Recasens, Vera Axelrod, and Jason Baldridge (2018). "Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns." In: *Transactions of the ACL*, to appear.

Welleck, Sean, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston (2019a). "Neural Text Generation with Unlikelihood Training." In: *CoRR* abs/1908.04319. arXiv: 1908.04319. URL: http://arxiv.org/abs/1908.04319.

Welleck, Sean, Jason Weston, Arthur Szlam, and Kyunghyun Cho (2019b). "Dialogue Natural Language Inference." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 3731–3741.

Williams, Adina, Nikita Nangia, and Samuel Bowman (2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122.

Wiseman, Sam, Alexander M. Rush, and Stuart Shieber (June 2016). "Antecedent Prediction Without a Pipeline." In: *CORBON*. San Diego, California, pp. 53–58. URL: https://aclweb.org/anthology/W16-0708.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing." In: *ArXiv* abs/1910.03771.

Wu, Jheng-Long and Wei-Yun Ma (2017). "A Deep Learning Framework for Coreference Resolution Based on Convolutional Neural Network." In: *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pp. 61–64.

Wu, Qianhui, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin (2020a). "Enhanced meta-learning for cross-lingual named entity recognition with minimal resources." In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9274–9281. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6466/6322.

Wu, Shijie and Mark Dredze (Nov. 2019). "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 833–

844. DOI: 10.18653/v1/D19-1077. URL: https://www.aclweb.org/anthology/D19-1077.

Wu, Wei, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li (July 2020b). "CorefQA: Coreference Resolution as Query-based Span Prediction." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6953–6963. DOI: 10.18653/v1/2020.acl-main.622. URL: https://www.aclweb.org/anthology/2020.acl-main.622.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." In: *CoRR* abs/1609.08144. arXiv: 1609.08144. URL: http://arxiv.org/abs/1609.08144.

Xia, Qingrong, Zhenghua Li, Min Zhang, Meishan Zhang, Guohong Fu, Rui Wang, and Luo Si (2019). "Syntax-aware neural semantic role labeling." In: *AAAI*, pp. 7305–7313.

Xing, Chen, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma (2017). "Topic Aware Neural Response Generation." In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 3351–3357.

Xu, Jingjing, Xuancheng Ren, Junyang Lin, and Xu Sun (2018). "Diversity-Promoting GAN: A Cross-Entropy Based Generative Adversarial Network for Diversified Text Generation." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3940–3949. DOI: 10.18653/v1/D18-1428. URL: https://www.aclweb.org/anthology/D18-1428.

Xu, Song, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou (July 2020). "Self-Attention Guided Copy Mechanism for Abstractive Summarization." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1355–1362. DOI: 10.18653/v1/2020.acl-main.125. URL: https://www.aclweb.org/anthology/2020.acl-main.125.

Xu, Wei, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch (2016). "Optimizing Statistical Machine Translation for Text Simplification." In: *Transactions of the Association for Computational Linguistics* 4, pp. 401–415.

Yang, Bishan and Tom Michael Mitchell (2017). "Leveraging Knowledge Bases in LSTMs for Improving Machine Reading." In: *ACL*.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le (2019). "XLNet: Generalized Au-

toregressive Pretraining for Language Understanding." In: *CoRR* abs/1906.08237.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann (2013). "WebAnno: A flexible, web-based and visually supported system for distributed annotations." In: *Proceedings of ACL*.

Yoon, Jaesik, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn (2018). "Bayesian Model-Agnostic Meta-Learning." In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31, pp. 7332–7342. URL: https://proceedings.neurips.cc/paper/2018/file/e1021d43911ca2c18-45910d84f40aeae-Paper.pdf.

Yu, Adams Wei, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le (2018). "QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension." In: *ICLR (Poster)*. OpenReview.net. URL: http://dblp.uni-trier.de/db/conf/iclr/iclr2018.html#YuDLZ00L18.

Zabell, Sandy L. (2005). *Symmetry and its discontents: essays on the history of inductive probability*. Cambridge University Press.

Zaheer, Manzil, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed (2020). "Big Bird: Transformers for Longer Sequences." In: *CoRR* abs/2007.14062. arXiv: 2007.14062. URL: https://arxiv.org/abs/2007.14062.

Zeman, Daniel et al. (2020). *Universal Dependencies 2.6*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. URL: http://hdl.handle.net/11234/1-3226.

Zeng, Xiangrong, Shizhu He, Kang Liu, and Jian Zhao (2018). "Large Scaled Relation Extraction With Reinforcement Learning." In: *AAAI*.

Zhang, Hongming, Yan Song, and Yangqiu Song (June 2019a). "Incorporating Context and External Knowledge for Pronoun Coreference Resolution." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 872–881. DOI: 10.18653/v1/N19-1093. URL: https://aclanthology.org/N19-1093.

Zhang, Hongming, Yan Song, and Yangqiu Song (June 2019b). "Incorporating Context and External Knowledge for Pronoun Coreference Resolution." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 872–881. DOI: 10.18653/v1/N19-1093. URL: https://www.aclweb.org/anthology/N19-1093.

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter J. Liu (2019c). *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*. arXiv: 1912.08777 [cs.CL].

Zhang, Sheng, Xin Zhang, Weiming Zhang, and Anders Søgaard (2020a). "Worst-Case-Aware Curriculum Learning for Zero and Few Shot Transfer." In: *arXiv preprint arXiv:2009.11138*. URL: https://arxiv.org/pdf/2009.11138.pdf.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020b). "BERTScore: Evaluating Text Generation with BERT." In: *Proceedings of the 8th International Conference on Learning Representations*. Virtual Conference, Formerly Addis Ababa Ethiopia.

Zhang, Weinan, Yue Zhang, Yuanxing Liu, Donglin Di, and Ting Liu (July 2019d). "A Neural Network Approach to Verb Phrase Ellipsis Resolution." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33, pp. 7468–7475. DOI: 10.1609/aaai.v33i01.33017468.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (2018). "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods." In: *NAACL*.

Zhou, Mingyuan, Lauren Hannah, David Dunson, and Lawrence Carin (2012). "Beta-Negative Binomial Process and Poisson Factor Analysis." In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, pp. 1462–1471. URL: http://proceedings.mlr.press/v22/zhou12c.html.

Zhou, Qingyu, Nan Yang, Furu Wei, and Ming Zhou (July 2017). "Selective Encoding for Abstractive Sentence Summarization." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1095–1104. DOI: 10.18653/v1/P17-1101. URL: https://www.aclweb.org/anthology/P17-1101.