Improving Generalization in Coreference Resolution via Adversarial Training

Sanjay Subramanian
University of Pennsylvania
sanjayssub34@gmail.com

Dan Roth
University of Pennsylvania
danroth@seas.upenn.edu

Abstract

In order for coreference resolution systems to be useful in practice, they must be able to generalize to new text. In this work, we demonstrate that the performance of the state-of-the-art system decreases when the names of PER and GPE named entities in the CoNLL dataset are changed to names that do not occur in the training set. We use the technique of adversarial gradient-based training to retrain the state-of-the-art system and demonstrate that the retrained system achieves higher performance on the CoNLL dataset (both with and without the change of named entities) and the GAP dataset.

Original: But **Dirk Van Dongen , president of the National Association of Wholesaler - Distributors** , said that last month 's rise " is n't as bad an omen " as the 0.9 % figure suggests . " If you examine the data carefully , the increase is concentrated in energy and motor vehicle prices , rather than being a broad - based advance in the prices of consumer and industrial goods , " **he** explained .
Replacement: Replace *Dick Van Dongen* with *Vendemiaire Van Korewdit*.

Table 1: An excerpt from the CoNLL test set. The coreference between the two highlighted mentions is correctly predicted by the (Lee et al., 2018) system, but after the specified replacement, the system incorrectly resolves "he" to a different name occurring outside this excerpt.

1 Introduction

Through the use of neural networks, performance on the task of coreference resolution has increased significantly over the last few years. Still, neural systems trained on the standard coreference dataset have issues with generalization, as shown by (Moosavi and Strube, 2018).

One way to improve the understanding of how a system overfits a dataset is to study the change in the system's performance when the dataset is modified slightly in a focused and relevant manner. We take this approach by modifying the test set so that each PER and GPE (person and geopolitical entity) named entity is different from those seen in training. In other words, we ensure that there is no leakage of PER and GPE named entities from the training set into the test set. We demonstrate that the performance of the (Lee et al., 2018) system, which is the current state-of-the-art, decreases when the named entities are replaced. An example of a replacement that causes the system to make an error is given in Table 1.

Motivated by these issues of generalization, this paper aims to improve the training process of neural coreference systems. Various regularization techniques have been proposed for improving the generalization capability of neural networks, including dropout (Srivastava et al., 2014) and adversarial training (Goodfellow et al., 2015; Miyato et al., 2017). The model of (Lee et al., 2018), like most neural approaches, uses dropout. In this work, we apply the adversarial fast-gradient-sign-method (FGSM) described by (Miyato et al., 2017) to the model of (Lee et al., 2018), and show that this technique improves the model's generalization even when applied on top of dropout.

The CoNLL-2012 Shared Task dataset (Pradhan et al., 2012) has been the standard dataset used for both training and evaluating English coreference systems since the dataset was introduced. The dataset includes seven genres that span multiple writing styles and multiple nationalities. We demonstrate that the system of (Lee et al., 2018) retrained with adversarial training achieves state-of-the-art performance on the original CoNLL-2012 dataset (Pradhan et al., 2012) as well as the CoNLL-2012 dataset with changed named entities. Furthermore, the system trained with the adversarial method ex-

hibits state-of-the-art performance on the GAP dataset (Webster et al., 2018), a recently released dataset focusing on resolving pronouns to people's names in excerpts from Wikipedia. The code and other relevant files for this project can be found via https://cogcomp.org/page/publication_view/871.

2 Related Work

(Moosavi and Strube, 2017, 2018) also study generalization of neural coreference resolvers. However, they focus on transfer and indicate that the ranking of coreference resolvers (trained on the CoNLL training set) induced by their performance on the CoNLL test set is not preserved when the systems are evaluated on a different dataset. They use the Wikicoref dataset (Ghaddar and Langlais, 2016), which is limited in that it consists of only 30 documents. They then show that the addition of features representing linguistic information improves the performance of a coreference resolver on the out-of-domain dataset.

The adversarial fast-gradient-sign-method (FGSM) was first introduced by (Goodfellow et al., 2015) and was applied to sentence classification tasks through word embeddings by (Miyato et al., 2017). Gradient-based adversarial attacks have since been used to train models for various NLP tasks, such as relation extraction (Wu et al., 2017) and joint entity and relation extraction (Bekoulis et al., 2018).

Our replacements of named entities can also be viewed as a way of generating adversarial examples for coreference systems; it is related to the earlier method proposed in (Khashabi et al., 2016) in the context of question answering and to (Alzantot et al., 2018), which provides a way of generating adversarial examples for simple classification tasks.

3 Adversarial Training for Coreference

In coreference resolution, the goal is to find and cluster phrases that refer to entities. We use the word "span" to mean a series of consecutive words. A span that refers to an entity is called a mention. If two mentions i and j refer to the same entity and mention i occurs before mention j in the text, we say that mention i is an antecedent of mention j. For a given mention i, the candidate antecedents of i are the mentions that occur before i in the text. In Figure 1, each line segment repre-
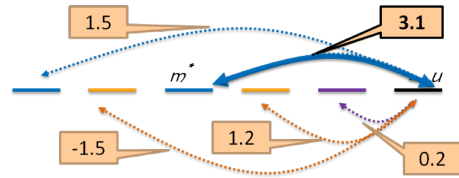

Figure 1: For each mention, the model computes scores for each of the candidate antecedent mentions and chooses the candidate with the highest score to be the predicted antecedent. This image was created by the authors of (Chang et al., 2013).

sents a mention and the arrows are directed from one mention to its possible antecedents.

We now review the model architecture of (Lee et al., 2018) and describe how we apply the fast-gradient-sign-method (FGSM) of (Miyato et al., 2017) to the model. Using GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018) embeddings of each word and using learned character embeddings, the model computes contextualized representations $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ of each word x_i in the input document using a bidirectional LSTM (Hochreiter and Schmidhuber, 1997). For candidate span i, which consists of the words at indices $start_i, start_i + 1, ..., end_i$, the model constructs a span representation $\mathbf{g}_i$ by concatenating $\mathbf{x}_{start_i}$, $\mathbf{x}_{end_i}$, $\frac{1}{\sum_{j=start_i}^{end_i} \beta_j} \sum_{j=start_i}^{end_i} \beta_j \mathbf{x}_j$, and $\phi(end_i - start_i)$, where the β_j's are learned scalar values and $\phi(\cdot)$ is a learned embedding representing the width of the span (Lee et al., 2017). The span representations are then used as inputs to feedforward networks that compute mention scores for each span and that compute antecedent scores for pairs of spans. In Figure 1, the number associated with each arrow is the antecedent score for the associated pair of mentions. The coreference score for the pair of spans (i, j) is the sum of the mention score for span i, the mention score for span j, and the antecedent score for (i, j). For each span i, the antecedent span predicted by the model is the span j that maximizes the antecedent score for (i, j). Let $\mathbf{g} = \{\mathbf{g}_i\}_{i=1}^{N}$ denote the set of the representations of all N candidate spans. Let $\mathcal{L}(\mathbf{g})$ denote the original model's loss function. (Note that the model's predictions and the loss depend on the input text only through the span representations.) For each $i \in \{1, ..., N\}$, let $\mathbf{g}_i^{adv}(\mathbf{g}) = \nabla_{\mathbf{g}_i} \mathcal{L}\left(\{\mathbf{g}_i\}_{i=1}^{N}\right)$ denote the gradient of the loss with respect to the span embeddings.

Then the adversarial loss with the FGSM is

$$\mathcal{L}_{adv}(\{\mathbf{g}_i\}_{i=1}^{N}) = \mathcal{L}\left(\left\{\mathbf{g}_i + \epsilon\frac{\mathbf{g}_i^{adv}(\mathbf{g})}{||\mathbf{g}_i^{adv}(\mathbf{g})||}\right\}_{i=1}^{N}\right).$$

The total loss used in training is

$$\mathcal{L}_{total}(\mathbf{g}) = \alpha\mathcal{L}(\mathbf{g}) + (1-\alpha)\mathcal{L}_{adv}(\mathbf{g}).$$

In our experiments, we find that $\alpha = 0.6$ and $\epsilon = 1$ work well. A key difference between our method and that employed by (Miyato et al., 2017) is that the latter applies the adversarial perturbation to the input embeddings, whereas we apply it to the span representations, which are an intermediate layer of the model. We found in our experiments that applying the FGSM to the character embeddings in the initial layer was not as effective as applying the method to the span representations as described above. Another difference between our method and that of (Miyato et al., 2017) is that we do not normalize the span embeddings before applying the adversarial perturbations.

4   No Leakage of Named Entities

Named entities are an important subset of the entities a coreference system is tasked with discovering. (Agarwal et al., 2018) provide the percentages of clusters in the CoNLL dataset represented by the PER, ORG, GPE, and DATE named entity types – 15%, 11%, 11%, and 4%, respectively. It is important for generalization that systems perform well with names that are different from those seen in training. We found that in the CoNLL dataset, roughly 34% of the PER and GPE named entities that are the head of a mention of some gold cluster in the test set are also the head of a mention of a gold cluster in the train set. Therefore, there is considerable overlap, or leakage, between the names in the train and test sets. In this section, we describe a method for evaluating on the CoNLL test set without leaked name entities.
We focus on PER and GPE named entities because they are two of the three most common entity types and because in general when replacing a PER or GPE name with another name, it is easy to not change the true coreference structure of the document. In particular, changing the name of an organization while ensuring that it is compatible with nominals in the cluster is nontrivial without a finer semantic typing. By contrast, we describe below how we control for gender and location type

when replacing PER and GPE names, respectively. We also ensure that the capitalization of the first letter in the replacement name is the same as in the original text. Finally, we note that the diversity of PER and GPE entities exceeds that of other named entity types; this increases the importance of generalization to new names and, at the same time, enables us to find matching names to use as replacements. Table 2 provides examples of text in the original CoNLL-2012 dataset and the corresponding text after our modifications.

4.1   Replacing PER entities

For replacing PER entities, we utilize the publicly available list of last names from the 1990 U.S. Census and a gazetteer of first names that has the proportion of people with this name who are males. The gazetteer was collected in an unsupervised fashion from Wikipedia. We denote the list of last names by $\mathcal{L}$, the list of male first names (i.e. first names with male proportion greater than or equal to 0.5 in the gazetteer) by $\mathcal{M}$, and the list of female first names (i.e. first names with male proportion less than or equal to 0.5 in the gazetteer) by $\mathcal{F}$. We remove all names occurring in training from $\mathcal{L}$, $\mathcal{M}$, and $\mathcal{F}$. We use the spaCy dependency parser (Honnibal and Johnson, 2015) to find the heads of each mention. We say that a mention is a person-mention if the head of the mention is a PER named entity, and we say that the name of the person-mention is the PER named entity that is its head. We use the dependency parser and the gold NER to identify all of the person-mentions. For each gold cluster containing a person-mention, we find the longest name among the names of all of the person-mentions in the cluster. If the longest name of a cluster has only one token, we assume that the name is a last name, and we replace the name with a name chosen uniformly at random from the remaining last names in $\mathcal{L}$. Otherwise, if the longest name has multiple tokens, we say that the cluster is male if the cluster contains no female pronouns ("she", "her", "hers") and one of the following is true: the first token does not appear in $\mathcal{M}$ or $\mathcal{F}$, if the token appears in $\mathcal{M}$, or the cluster contains a male pronoun ("he", "him", "his"). We say that the cluster is female if it is not male. Then we (1) replace the last token with a name chosen uniformly at random from the remaining last names in $\mathcal{L}$, and (2) replace the first token with a name chosen uniformly at random from the remaining

Original	No Leakage
We asked **Judy Muller** if she would like to do the story of a fascinating man . She took a deep breath and said , okay .	We asked **Sallie Kousonsavath** if she would like to do the story of a fascinating man . She took a deep breath and said , okay .
The last thing President **Clinton** did today before heading to the Mideast is go to church – appropriate , perhaps , given the enormity of the task he and his national security team face in the days ahead .	The last thing President **Golia** did today before heading to the Mideast is go to church – appropriate , perhaps , given the enormity of the task he and his national security team face in the days ahead .
In theory at least , tight supplies next spring could leave the wheat futures market susceptible to a supply - demand squeeze , said Daniel Basse , a futures analyst with AgResource Co. in **Chicago** .	In theory at least , tight supplies next spring could leave the wheat futures market susceptible to a supply - demand squeeze , said Daniel Basse , a futures analyst with AgResource Co. in **Machete** .

Table 2: Excerpts from the CoNLL-2012 test set and their versions after we have replaced PER and GPE names to avoid name leakage.

first names in $\mathcal{M}$ if the cluster is male or from the remaining first names $\mathcal{F}$ if the cluster is female. Note that our sampling from each of $\mathcal{L}$, $\mathcal{M}$, and $\mathcal{F}$ is without replacement, so no last name is used as a replacement more than once, no male first name is used more than once, and no female first name is used more than once.

4.2 Replacing GPE entities

Our approach to replacing GPE entity names is very similar to that used for PER names. We use the GeoNames[1] database of geopolitical names. In addition to providing a list of GPE names, this database also categorizes the names by the type of entity to which they refer (e.g. city, state, county, etc.). The data includes the names and categories of more than $11,000,000$ locations in the world. We restrict our attention to GPE entities that satisfy the following requirements: (1) they occur in the GeoNames database and (2) they are not countries. We say that a mention is a GPE-mention if its head (as given by the dependency parser) is a GPE named entity satisfying these three requirements. (Again, we use the gold NER to identify GPE names in the CoNLL text.) We remove all GPE names occurring in the training set from the list of replacement GPE names for each location category. Then for each cluster containing a GPE-mention, we find the GeoNames category for the mention's GPE name and replace the name with a randomly chosen name from the same category. As with PER names, we sample names from each

category without replacement, so each GPE name is used for replacement at most once.

5 Experiments

We trained the (Lee et al., 2018) model architecture with the adversarial approach on the CoNLL training set for 355000 iterations (the same number of iterations for which the original model was trained) with the same training hyperparameters used by original model. For comparing with the (Lee et al., 2017) and (Lee et al., 2018) systems, we use the pretrained models released by the authors.[2]

The datasets used for evaluation are the CoNLL and GAP datasets.

5.1 CoNLL Dataset

Table 3 shows the performance on the CoNLL test set, as measured by CoNLL F1, of the (Lee et al., 2018) system with and without our adversarial training approach. The replacement of PER and GPE entities decreased the performance of the original system by more than 1 F1.

5.2 GAP Dataset

The GAP dataset (Webster et al., 2018) focuses on resolving pronouns to named people in excerpts from Wikipedia. The dataset, which is gender-balanced, consists of examples in which

---

[1]http://www.geonames.org/

	Original	No Leakage
(Lee et al., 2018)	72.96	71.86
+Adv. Training	**73.23**	**72.36**

Table 3: Results (CoNLL F1) on the CoNLL Test Set. "Original" refers to the original test set, and "No Leakage" refers to the test set modified with the replacement of named entities described in Section 4. For each dataset, highest score for each dataset is **bolded** and is underlined if the difference between it and the other model's score is statistically significant ($p < 0.20$ per a stratified approximate randomization test similar to that of (Noreen, 1989)).

	M	F	O
(Lee et al., 2017)	68.7	60.0	64.5
(Lee et al., 2018)	75.8	70.6	73.3
+Adv. Training	**77.3**	**72.1**	**74.7**

Table 4: Results (F1 metric defined by (Webster et al., 2018)) on the GAP Test Set. **M** refers to male pronouns, **F** refers to female pronouns, and **O** refers to the full evaluation data. For each category, highest score is **bolded** and underlined if difference between it and next-highest score is statistically significant ($p < 0.05$ per the McNemar test (McNemar, 1947)).

the system must determine whether a given pronoun refers to one, both, or neither of two given names. Thus, the task can be viewed a binary classification task in which the input is a (pronoun, name) pair and the output is True if the pair is coreferent and False otherwise. Performance is evaluated using the F1 score in this binary classification setup. Table 4 shows the performance on the GAP test set of the (Lee et al., 2017)[3] and (Lee et al., 2018) systems as well as the system trained with our adversarial method. The adversarially trained system performs significantly better over the entire dataset in comparison to the previous systems, and the difference is consistent between genders. In particular, we observe that the bias (i.e. ratio of female to male F1 score) is roughly the same (0.93) for the (Lee et al., 2018) system with and without adversarial training and that this bias is better (i.e. the ratio is closer to 1) than that exhibited by the (Lee et al., 2017) system (0.87).

---

[3]The results that we report for the (Lee et al., 2017) system differ slightly from those reported in Table 10 of (Webster et al., 2018) due to a difference in the parser and potentially small differences in the algorithm for converting the system's output to the binary predictions necessary for the GAP scorer.

6 Conclusion

We show that the performance of the (Lee et al., 2018) system decreases when the names of PER and GPE entities are changed in the CoNLL test set so that no names from the training set leak to the test set. We then retrain the same system using an application of the fast-gradient-sign-method (FGSM) of adversarial training, showing that the retrained system consistently performs better on the original CoNLL test set, the CoNLL test set with No Leakage, and the GAP test set. Our new model is a new state-of-the-art for all these data sets.

Acknowledgements

References

Oshin Agarwal, Sanjay Subramanian, Ani Nenkova, and Dan Roth. 2018. Named person coreference in english news. *arXiv preprint arXiv:1810.11476*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836.

Kai-Wei Chang, Rajhans Samdani, and Dan Roth. 2013. A constrained latent variable model for coreference resolution. In *EMNLP*.

Abbas Ghaddar and Philippe Langlais. 2016. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *LREC*.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question answering via integer programming over semistructured knowledge. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. 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)*, pages 687–692.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2017. Adversarial training methods for semisupervised text classification. *ICLR*.

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19.

Nafise Sadat Moosavi and Michael Strube. 2018. Using linguistic features to improve the generalization capability of neural coreference resolvers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, 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)*, pages 2227–2237.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguou. In *Transactions of the ACL*, page to appear.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783.