Guy or Hombre? Data Augmentation in Conversational Coreference Resolution

Fran Pugelnik, Sara Borzić, Nera Frajlić

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{fran.pugelnik, sara.borzic, nera.frajlic}@fer.com

Abstract

Coreference resolution is a challenging task in natural language processing and requires a specifically labeled data set. Labeled textual data is a luxury and often requires human interference. Therefore, instead of collecting more data, proper data augmentation can improve model performance. We test the performance of AllenNLP's coreference resolution model using augmented transcripts from the TV show *Friends*. The idea is to construct additional training instances using synonym replacement. We examine if this augmentation method has a desired boost in performance for coreference resolution task.

1. Introduction

The amount and quality of data greatly affects the performance of any model in any NLP task, including coreference resolution. Instead of collecting more data, it might be more convenient to apply data augmentation to already obtained data to generate additional, synthetic data and make the model generalize better. In this paper, we describe an application of simple data augmentation techniques for improving performance on coreference resolution task. The model chosen for tackling the coreference resolution task is an end-to-end neural model by Lee et al. (2017). Its key idea is to consider all spans in a document as potential mentions to later produce the most likely correct clustering for mentions.

The advantage of using a neural model for coreference resolution is the usage of word embeddings to capture the similarity between words. This can lead to the prediction of false-positive links when the model conflates paraphrasing with kinship or similarity (Lee et al., 2017). We dive deeper into this problem in Section 4. of this paper.

2. Related Work

Manipulations in the input data in NLP tasks often result in system failure, although they would not affect human performance on the same task. Our idea is based on many recent papers that expose the benefits of data augmentation techniques in such cases, and for natural language processing tasks in general. A lot of existing work focuses on using data augmentation for mitigating gender bias in NLP tasks (Zmigrod et al., 2019; Zhao et al., 2018; Lu et al., 2018).

Some similar approaches were explored for improving syntactic parsing (Elkahky et al., 2018), as well as natural language inference (NLI) (Min et al., 2020) and NLI and sentiment analysis (Kaushik et al., 2020). Contextual data augmentation is described in (Wu et al., 2018; Kobayashi, 2018) on various text classification tasks, and other data augmentation operations such as synonym replacement, random insertion, random swap and random deletion are used in (Wei and Zou, 2019) for text classification tasks.

Other interesting techniques, such as back-translation and word replacing with TF-IDF are discussed in (Xie et al., 2019) also for text classification tasks. Augmentation of a word by a contextual mixture of multiple related words is used in (Gao et al., 2019) for the task of machine translation. (Wu et al., 2019) explored the use of existing question answering datasets as data augmentation for coreference resolution. For the task of named entity recognition, various approaches such as label-wise token replacement, synonym replacement, mention replacement, and shuffle within segments are described in (Dai and Adel, 2020). The most natural choice for data augmentation – replacing the words with their synonyms is also used in (Zhang et al., 2015) for controlling generalization error.

Neural models often perform well by using superficial features, rather than more general ones that are preferred and more natural to humans. Data augmentation is used in (Jha et al., 2020) by generating training examples to encourage the model to focus on the strong features.

Considering the previous success of data augmentation techniques on a broad spectrum of topics we think that it might be useful to apply synonym-based augmentation on coreference resolution task in conversational text.

3. Database

The data collection created by Chen and Choi (2016) consists of transcripts of the first two seasons from the TV show *Friends*[1]. Most of the dialogues between the characters are everyday conversations and introduce over 200 speakers. Figure 1 shows an example of a multiparty dialogue present in the dataset. The mention "mom" is not one of the speakers; nonetheless, it refers to the specific person, Judy, that could appear in some other dialogue. Identifying and clustering such mentions might require cross-document coreference resolution.

The dataset follows the CoNLL 2012 Shared Task data format[2]. The data we used consists of documents, each document is delimited and each episode is considered a document. Each word is associated with a word form, part-of-speech tag, and its lemma. A speaker is annotated for each sentence in the training set, as well as in the test set. Each mention is annotated with a belonging entity ID that is con-

[1] `https://bit.ly/2SXXn7q`

[2] `https://bit.ly/3j1mUau`

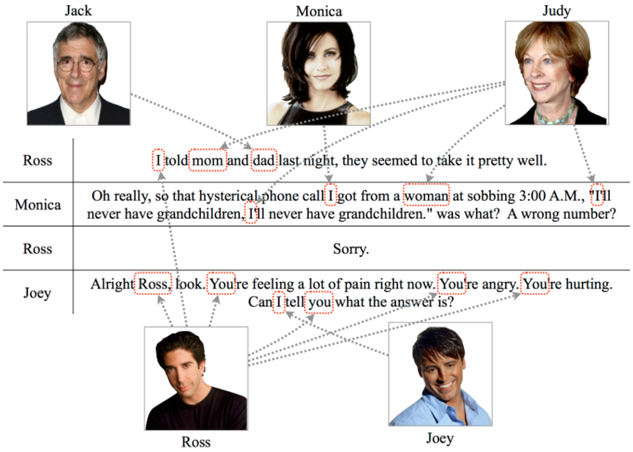
Jack
Monica
Judy
Ross
I told mom and dad last night, they seemed to take it pretty well.
Monica
Oh really, so that hysterical phone call I got from a woman at sobbing 3:00 A.M., "I'll never have grandchildren, I'll never have grandchildren." was what? A wrong number?
Ross
Sorry.
Joey
Alright Ross, look. You're feeling a lot of pain right now. You're angry. You're hurting. Can I tell you what the answer is?
Ross
Joey

Figure 1: Dialogue example with labeled mentions assigned to entities. Taken from `https://competitions.codalab.org/competitions/17310`

sistent across all documents. For example, Table 1[3] shows a specific sentence from the dataset. The sentence is from the second episode in season 2, the scene number is 0 and the speaker is Chandler Bing. The entity of the mention in this example belongs to their neighbour, the character Ugly Naked Guy.

4. Our Approach

Our experiment consists of an end-to-end neural model by Lee et al. (2017) which will be trained on a semeval 2018 dataset and its augmented counterpart.

4.1. Model's Weaknesses

In the model we used, "The key idea is to directly consider all spans in a document as potential mentions and learn distributions over possible antecedents for each" (Lee et al., 2017). The spans that are not likely to appear in the clusters are discarted. For the spans remaining after this step the model decides whether the span is coreferent with some of the antecedent spans. It returns the resulting coreference links which (after applying transitivity) imply clustering of spans for the given document.

This model uses word embeddings to capture the similarity between words. This can lead to a prediction of false positives when the model confuses paraphrasing with relatedness or similarity. The example given to illustrate this scenario was when the model falsely predicted a link between a pilot and a flight attendant or Prince Charles and his wife Camilla and Charles and Diana.

Another concern is that the model sometimes confuses in rarely occurring patterns and examples that humans wouldn't find challenging.

4.2. Idea

It is assumed that both of the problems explained in 4.1. could be mitigated with a larger corpus of training data which would overcome the sparsity of these patterns.

We argue that using data augmentation to extend our corpus would help the model generalize and improve its results on patterns like these while it wouldn't affect the rest of its performance.

For creating a larger dataset we augmented certain words from the existing data set and appended it to the original, which is described in Section 4.3.. One can say that by augmenting the training data with synonyms, we would create duplicate data. However, that is not the case because synonyms from WordNet often don't match the exact synonyms of words from the dataset. As an example, the word 'date', which meant an appointment to meet someone, was in one case replaced with the word 'engagement' which is arguably not a synonym in this context. This is how the noise is added to our corpus and how the better generalization of the model is accomplished.

4.3. Data Augmentation

Our approach consists of using data augmentation techniques for improving AllenNLP's model's performance on coreference resolution task. *nlpaug*[4] is a library for textual augmentation in machine learning experiments. The experiment consists of using the augmenter from WordNet[5] lexical database, which replaces words from the input dataset with its synonyms. More specifically, lemmas of words were replaced with their synonyms. An example of different sentences generated from the same sentence using synonym augmentation is presented in Table 2.

Lemmas that were excluded from the synonym replacement task are stopwords from Nltk[6] (Natural language toolkit). Additionally, to assure that words that greatly affect our task of coreference resolution stay the same, entities (mostly names of the characters) were also excluded from synonym replacement operation. Due to limitations imposed by the .conll file format, it is decided that if a particular synonym does not consist of one word, but is rather a phrase, it will not be a candidate for replacement. The words whose lemmas were replaced by synonyms, had their *Word form* also replaced with the same lemma form. Having that in mind, some information in the created dataset was lost. However, The POS and constituency tags from .conll file were not changed so the replacement of the *Word form* shouldn't cause a lack of too much information. After creating such augmentation, it was simply appended to the existing dataset and presented to the model.

5. Experimental Setup

The pretrained model in AllenNLP library was trained on the English coreference resolution data from the CoNLL-2012 shared task (Pradhan et al., 2012). We tested the model's performance on data described in Section 3., as well as the augmented data described in Section 4.3..

Inspired by Dai and Adel (2020), we simulate a low-resource setting and select the first 16 and 32 episodes from the training set to create the corresponding small and medium training sets to perform data augmentation and ob-

[3] Some columns were omitted from this preview for clarity

[4] `https://nlpaug.readthedocs.io/en/latest/`

[5] `https://wordnet.princeton.edu/`

[6] `https://www.nltk.org/`

Table 1: Reduced training dataset example.

Document ID	Scene ID	Token ID	Word form	POS tag	Lemma	Speaker	Entity ID
/friends-s01e02	0	0	Ugly	JJ	ugly	Chandler_Bing	(380
/friends-s01e02	0	1	Naked	JJ	naked	Chandler_Bing	-
/friends-s01e02	0	2	Guy	NNP	guy	Chandler_Bing	380)
/friends-s01e02	0	3	got	VBD	get	Chandler_Bing	-
/friends-s01e02	0	4	a	DT	a	Chandler_Bing	-
/friends-s01e02	0	5	Thighmaster	NN	thighmaster	Chandler_Bing	-
/friends-s01e02	0	6	!	.	!	Chandler_Bing	-

Table 2: Data augmentation examples.

ORIG.	Sounds like a date to me.
AUGM.	Sound like a particular date to me. Speech sound like a engagement to me. Sounds comparable a engagement to me.

serve the results. We expect to get the best results by training the model on the largest train set.

For each training set (small, medium, complete) we conduct simple experiments. We split the training set using random seeds on training and validation sets (87.5% - 12.5%). The reason for choosing the unusual train-dev split sizes is because, in this way, an integer division of episodes in the training set is ensured. We then apply data augmentation on the training set and test the model performance on the test set (which is always the same). This experiment is repeated 5 times for each training set size using different random seeds so statistical evaluations could be performed.

We then augmented each of the training sets by adding synonym equivalents for 50% of the training set. The final result were 6 training sets - small, medium, complete, and their corresponding augmented sets.

6. Results

The results are presented in Table 3. The table shows macro-averaged metrics (precision, recall, and F1-score) obtained from testing the model trained on a specific training set (S - small, M - medium, F - full/complete).

6.1. Analysis

As we were unable to perform testing on a large number of training instances (5 runs with different random seeds) due to the time-complexity of the training process, we can't say much about the distribution of test results. However, the observation distribution pairs (e.g. small original - small augmented) are similar in shape, so we ran a non-parametric Mann-Whitney U test in place of an unpaired t-test. We wanted to test whether the observations (specifically F1-scores) in one sample tend to be larger than observations in the other. Considering the fact that the mean results shown in Table 3 fluctuate in different ways, we state specific research hypotheses for each training set size and perform a test calculation at a significance level of 0.05.

For both small and medium data set, the research hypothesis we chose to state is that a randomly selected F1-score obtained from the population of tests conducted on non-augmented data is greater than the same score obtained when testing on augmented data. Informally, after seeing the results presented in Table 3, we decided to argue that the results are significantly better when using non-augmented data. The null hypothesis is rejected in favor of the alternative. Therefore, it appears that this type of data augmentation impairs this models performance.

As for the full data set, according to the mean values of F1-scores, we chose to test the hypothesis that a randomly selected F1-score obtained from the population of tests conducted on non-augmented data differs in any way from F1-score obtained when testing on augmented data. The resulting p-value equals 0.15, therefore the null hypothesis cannot be rejected. We can conclude that no significant difference can be confirmed between the distributions of F1-scores from testing the model on the full data set and the full augmented data set.

6.2. Discussion

The results show that data augmentation with synonyms doesn't improve the performance of the coreference resolution model. Moreover, all subsets of the augmented datasets show lower performance on the test set than the corresponding subsets of non-augmented data. We argue that this is the result of the generalization problem presented in Section 4.1.. Obviously, the polysemy of words has created more problems than benefits for the model. A possible improvement is to use contextual word embeddings in combination with WordNet synonyms for augmenting data so the error due to the lack of context lapses. Another idea would be to include word or span representations that can distinguish between equivalence and paraphrasing. Moreover, some generalization techniques should be applied. In terms of other augmentation methods, antonym replacement or random word generation might help with generalization.

Secondly, we can see that the obtained results have consistently shown that the model achieves a lot lower recall than precision. Knowing that recall denotes how often our model has classified data correctly in regards to the full set of relevant results, we can notice that the model has more problems with recognizing phrases that refer to some entity than it does with correctly deciding on which entity that

Table 3: Results.

	S			M			F		
Data	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
original	0.564	0.431	**0.480**	0.581	0.396	**0.453**	0.579	0.318	0.371
synonym augmented	0.546	0.327	0.395	0.553	0.301	0.356	0.581	0.224	0.285

phrase is referring to. Therefore, it might be interesting to look into the reasons for such phenomena in future work.

7. Conclusion

We presented a simple data augmentation technique for tackling the task of coreference resolution in multiparty dialogues. We showed that synonym replacement for this specific data and model showed no statistically significant improvement in regards to non-augmented data.

Coreference resolution is a challenging NLP problem. The model had problems detecting mentions which can be seen from the low recall scores. While precision scores are better, this model wouldn't be useful in real-world application and human performance on this task is still significantly better.

References

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, Los Angeles, September. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas, November. Association for Computational Linguistics.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. *CoRR*, abs/2010.11683.

Ali Elkahky, Kellie Webster, Daniel Andor, and Emily Pitler. 2018. A challenge set and methods for noun-verb ambiguity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2572, Brussels, Belgium, October-November. Association for Computational Linguistics.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544, Florence, Italy, July. Association for Computational Linguistics.

Rohan Jha, Charles Lovering, and Ellie Pavlick. 2020. When does data augmentation help generalization in nlp? *CoRR*, abs/2004.15012.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. 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 452–457, New Orleans, Louisiana, June. Association for Computational Linguistics.

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, Copenhagen, Denmark, September. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *CoRR*, abs/1807.11714.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online, July. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July. Association for Computational Linguistics.

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, Jeju Island, Korea, July. Association for Computational Linguistics.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, abs/1901.11196.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2018. Conditional BERT contextual augmentation. *CoRR*, abs/1812.06705.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2019. Coreference resolution as query-based span prediction. *CoRR*, abs/1911.01746.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation. *CoRR*, abs/1904.12848.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna M. Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *CoRR*, abs/1906.04571.