Evaluation of Coreference Resolution Systems Under Adversarial Attacks

Haixia Chai[†] **Wei Zhao**[Φ] **Steffen Eger**[Φ] **Michael Strube**[†]
[†] Heidelberg Institute for Theoretical Studies
[Φ] Computer Science Department, Technische Universität Darmstadt
{haixia.chai, michael.strube}@h-its.org
{zhao,eger}@aiphes.tu-darmstadt.de

Abstract

A substantial overlap of coreferent mentions in the CoNLL dataset magnifies the recent progress on coreference resolution. This is because the CoNLL benchmark fails to evaluate the ability of coreference resolvers that requires linking novel mentions unseen at train time. In this work, we create a new dataset based on CoNLL, which largely decreases mention overlaps in the entire dataset and exposes the limitations of published resolvers on two aspects—lexical inference ability and understanding of low-level orthographic noise. Our findings show (1) the requirements for embeddings, used in resolvers, and for coreference resolutions are, by design, in conflict and (2) adversarial approaches are sometimes not legitimate to mitigate the obstacles, as they may falsely introduce mention overlaps in adversarial training and test sets, thus inflating the performance.

1 Introduction

Resolution of coreferring expressions is a natural step for text understanding, but coreference resolvers appear to have a negligible effect in downstream NLP tasks (Yu and Ji, 2016; Durrett et al., 2016; Voita et al., 2018). For instance, Durrett et al. (2016) rewrite pronouns with their antecedents (e.g., *he* is replaced by *Dominick Dunne*), using the Berkeley Entity Resolution System (Durrett and Klein, 2014). However, this fails to improve the cross-sentence coherence of system summaries, although the resolver performs well on the OntoNotes 4.0 dataset (Pradhan et al., 2011).

The CoNLL benchmark (Pradhan et al., 2012) reflects the recent advances of coreference resolution systems. Nevertheless, previous work (Moosavi and Strube, 2017) indicates that the progress on the CoNLL benchmark is inflated, as the training and test sets share a large size of mentions. This may be the reason why coreference resolvers have little effect in downstream tasks.

Test Example: Iraqi leader Saddam has given a speech to mark the tenth anniversary of **the Gulf war**. The Iraqi leader said **the Gulf war** was a confrontation...
Train Example: There were other signs today that Iraq's leaders have few regrets over the action that precipitated **the Gulf war**. **The Gulf war** began 10 years ago...

Table 1: Replacing "*the Gulf war*" with "*the Gulf warfare*" or "*the Gulf wärfäre*" addresses (1) exact match in the test example; (2) mention overlaps across examples.

As opposed to evaluating on standard benchmarks, recent work (Glockner et al., 2018; Pruthi et al., 2019; Eger et al., 2019; Eger and Benz, 2020) investigates the generalization ability of NLP systems under adversarial attacks. For instance, Glockner et al. (2018) show that natural language inference systems fail blatantly when lexical changes, e.g., replacing a word by its synonym, occur in premises and hypotheses. Pruthi et al. (2019) observe that spelling errors distract text classification systems from correct prediction. Inspired by these works, we investigate published coreference resolvers in two realistic adversarial setups, which challenge (a) lexical inference ability to resolve coreferent mentions, where one mention is, e.g., synonymous or in a type-of relationship with its antecedent and (b) denoising ability against typographic (low-level) noise. To do so, we construct a new benchmark dataset by modifying the mention spans from CoNLL (Pradhan et al., 2012). This can mitigate lexical overlaps between the CoNLL training and test sets, as illustrated in Table 1.

Our analysis yields several findings: (1) We show that the lexical inference ability of published resolvers, including the state-of-the-art resolver based on BERT, is poor, i.e., the failure to properly resolve the coreference of a mention and its hy-

Proceedings of the First Workshop on Computational Approaches to Discourse, pages 154–159

https://doi.org/10.18653/v1/P17

pernymous (or hyponymous) antecedent within the same synset. (2) We identify an important reason for this failure: a mismatch, by design, between the requirements of coreference resolution and embeddings (used in resolvers). While a plausible coreference resolver anticipates ignoring the semantic difference of a word and its hypernym and linking them as coreferent mentions, embeddings capture the nuanced and fine-grained meanings well. (3) Further, we show that coreference resolvers fail to generalize to the CoNLL benchmark dataset with minor low-level (orthographic) noise. As a remedy, we use a common adversarial approach (Goodfellow et al., 2015) to incorporate lexical changes and low-level noise in coreferent mentions at train time, which appears to largely address the obstacles. However, we reveal that it introduces a large size of mention overlaps in the adversarial training and the test sets. This indicates an unrealistic situation where resolvers are only robust to what has been seen during training.

These findings indicate potential directions for future work, which may benefit coreference resolvers in downstream tasks and in real-world applications with natural occurring noise (e.g., user-generated texts).

2 Adversarial Data Collection

Our goal is to construct a benchmark dataset on which we evaluate the ability to resolve coreference that requires lexical inference and understanding of low-level noise.

2.1 Generating Adversarial Examples

Recent work for adversarial attacks concerning lexical changes and orthographic modification has shown deficiencies of NLP models for many tasks. To adapt previous approaches to coreference resolution, we design the following attack schemes where we focus on text changes occurring in mention spans. This setup also can address lexical overlap issue. To do so, we collect mentions from the training and test sets in the CoNLL benchmark dataset. We i.i.d. randomly attack each word in a mention with probability p and apply one of the below schemes. Table 2 shows examples of our modifications.

Lexical Changes. Modifiers and head words of noun phrases in a chain of mentions sometimes occur repeatedly. For instance, *president* both appears in the mention *the 44th president of the US* and its antecedent *the first African-American president of the US*. The CoNLL dataset involves many such lexical overlaps in coreferent mentions. Furthermore, Moosavi and Strube (2017) find a large size of mentions are overlapping in the CoNLL training and test examples. Together, this shows that the CoNLL evaluation setup does require only little lexical inference requirements. Subramanian and Roth (2019) remove named entities overlapping in the training and test sets. In contrast, we choose a word overlap randomly from mentions and substitute it with its hyponym, hypernym and synonym, as found in WordNet (Miller, 1995). To prevent the meaning of a word substitution deviated from the original word, we make the substitution only when two words share one word sense (synset), obtained from adapted LESK algorithm (Banerjee and Pedersen, 2002).

Modification	Original → Modification
SWAP	people → peolpe
DELETE	rise → rse
VISUAL	emergency → emergeňcy
SYNONYM	next → upcoming
HYPONYM	people → workers
HYPERNYM	pigeon → bird

Table 2: Examples of text modification.

Orthographic Changes. Character-level ("low-level") text changes, e.g., random swapping of characters (Pruthi et al., 2019), create surface form noise that often does not affect humans. We investigate the impact of different forms of low-level noise, namely (a) swapping a pair of adjacent letters, (b) deleting letters, and (c) visual perturbation, i.e., changing characters in a word by visually similar ones. To make text changes less perceptible to humans, we restrict for (a) and (b) to: (1) an individual word is allowed to be modified only once, (2) the first and the last letter of a word cannot be modified—as human reading is more resilient to internal letter exchanges, as shown by psycholinguistic research (Davis, 2003), and (3) modifications to a word with less than four characters are not allowed. As for visual attacks (c), we obtain character 'embeddings' from descriptions of each character in the Unicode 11.0.0 final names list, and then determine a set of nearest neighbors by choosing those characters whose descriptions refer to the same letter. Such perturbations have been shown little effect on human text processing (Eger et al., 2019).

Systems	CLEAN	Avg	△	$\alpha(3)$	△	$\beta(3)$	△
Non-Neural Systems							
DETERMINISTIC	57.10	46.32	−10.78	41.24	−15.86	51.40	−5.70
STATISTICAL	66.83	55.17	−11.66	50.24	−16.59	60.10	−6.73
Neural Systems							
DEEP-RL	69.13	58.15	−10.98	51.17	−17.96	65.12	−4.01
COARSE-TO-FINE (C2F)	72.96	60.04	−12.92	55.08	−17.33	64.99	−7.97
C2F⊕BERT	73.38	61.59	−11.79	55.63	−17.75	67.54	−5.84
C2F⊕SPANBERT	77.43	64.62	−12.81	58.44	−18.99	70.80	−6.63

Table 3: Overall results of the published baselines, on the clean, α (orthographic noise) and β (lexical changes) test sets. Brackets denote the number of modified test sets per group (α or β). Results are averaged for each group. △ is the difference between the performance of the clean and average result per group.

3 Experiments

Benchmark Dataset. We collect the training, development and test documents in the CoNLL benchmark dataset and use the above-described adversarial schemes to generate 16,812 training, 2,058 development and 2,088 test documents. We note that there are only about 2.3 words per mention and about 2 mentions per sentence on average in the CoNLL dataset. Therefore, we set a relatively low modification probability $p = 0.5$, thus making about 2 words changes per sentence. The percents of the mentions in the CoNLL test set modified by lexical and orthographic changes are 24% and 46%, respectively. When applying text changes to the test set, the percent of mention overlaps in the training and the test sets are decreased from 56.7% to 34.3%.

Baselines. We investigate non-neural systems[1], namely the DETERMINISTIC (Lee et al., 2013) and STATISTICAL (Clark and Manning, 2015) systems together with neural systems, including DEEP-RL (Clark and Manning, 2016), COARSE-TO-FINE (C2F) (Lee et al., 2018), C2F⊕BERT and C2F⊕SPANBERT (Joshi et al., 2019). The results are reported using the CoNLL F1 score—the average of MUC (Vilain et al., 1995), B3 (Bagga and Baldwin, 1998) and CEAFe (Luo, 2005).

Overall Results. Despite the minor changes in text, Table 3 shows that, the drop in performance is consistently big on average (10-12 points CoNLL F-score) across systems. The systems appear to suffer the most from orthographic changes, however, the percent of the examples of low-level noises is twice as large as that of lexical changes. Together, this exposes the limitation of non-neural and neural systems, including the systems based on BERT and SpanBERT, on lexical inference ability and understanding of low-level noise. Also, we note that the drop in non-neural baselines is smaller, which we believe is because linguistic features are primary predictors in them and have a positive effect.

[1]For non-neual systems, their linguistic features are extracted from our benchmark dataset using spaCy.

Training Set	CLEAN	SYNO	HYPO	HYPER
100% CLEAN	73.4	69.1	67.9	65.6
50% CLEAN and 50% SYNONYM	72.7	71.8	70.4	69.2

Table 4: Results of C2F⊕BERT on the test sets.

4 Shielding via Adversarial Training

Shielding Setup. We measure to what extent adversarial training (Goodfellow et al., 2015) can improve lexical inference ability and the robustness to low-level noise for the baseline systems. We include the adversarial training set at train time, but do not augment the training data, i.e., only replace 50% clean examples using our text manipulations. We split our evaluation into two setups: (1) in-domain evaluation, e.g., the training and test set used for training and evaluation are modified by swapping characters and (2) out-of-domain evaluation, e.g., we use adversarial training that trains a baseline system from scratch on a modified training set of one noise, denoted as AT-NOISE, and evaluates on the adversarial test sets of the remaining noise.

Lexical Changes Analysis. Table 4 shows that the performance drops for C2F⊕BERT in the HYPONYM and HYPERNYM test sets are much bigger than that in the SYNONYM test set, but AT-SYNONYM considerably helps. To more thoroughly examine this, we randomly extract pairs of 1,000

words and their synonyms, hyponyms and hypernyms from WordNet, as a form of coreferent mentions. We show histograms of the cosine similarity scores of word pairs, based on the last layer of BERT embeddings, used in C2F⊕BERT. Figure 1 (above) shows that a pair of a mention and its hypernymous/hyponymous antecedent is often assigned lower a cosine similarity score than a mention and its synonymous antecedent pair, suggesting that BERT embeddings capture the semantic differences of the three well. However, a plausible coreference resolver requires to ignore such fine-grained differences in meanings and links them all as coreferent mentions. This indicates the requirements for embeddings, used in resolvers, and for coreference resolvers, by design, are in conflict. However, this issue can be mitigated using AT-SYNONYM, as illustrated in Figure 1 (below). This is because a gold label can bridge a mention and its hypernymous/synonymous antecedent (within the same synset), thus omitting the semantic differences of them.

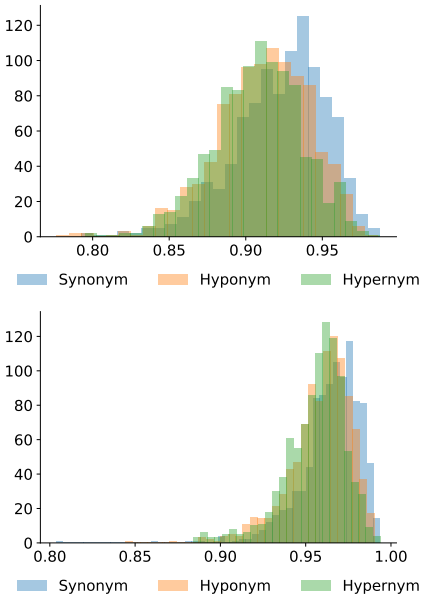

Figure 1: Histograms of cosine similarity scores of word pairs. C2F⊕BERT trained on the clean training set (above) and on SYNONYM training set (below).

In-domain and Out-of-domain Evaluations. Figure 2 shows that C2F⊕BERT via adversarial training appears to achieve consistent improvements in the in-domain evaluation setup, e.g., the gain achieved by AT-SWAP is 15.3 points on the SWAP test set. However, we observe that about 10% percent of mention are overlapping in the adversarial training and test sets, introduced by the

Training Set	SWAP	DELETE	VISUAL
100% CLEAN	56.8	55.4	54.5
100% SYNONYM	50.1	48.7	48.0
50% CLEAN and 50% SYNONYM	**58.1**	**57.1**	**55.6**

Table 5: Results of C2F⊕BERT trained via AT, on the training sets with synonym changes.

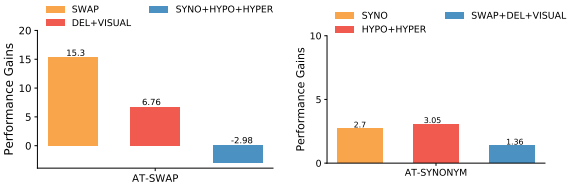

Figure 2: Performance gains (in points) in the in-domain and out-of-domain evaluation setup.

adversarial training approach. This may give a false and inflated impression for the improvements. Further, the effects for the out-of-domain evaluation are different. For instance, AT-SWAP obtains a large gain (+6.76 points) on the DELETE and VISUAL test sets, as the domain difference between the two and the SWAP test set is small. However, we note that AT-SWAP has a negative effect for the performance on the adversarial test sets involving lexical changes, since character-level noise and lexical replacement have little in common. In contrast, AT-SYNONYM appears to have a positive effect for the performance in the low-level noise domain. However, Table 5 shows that C2F⊕BERT trained on full SYNONYM training set causes a big performance drop on average across low-level noise. This indicates that enriching the system with lexical knowledge fails to improve its robustness to orthographic changes (similarly as for the negative effect of AT-SWAP to lexical changes). The gain on the test sets with low-level noise only appears when involving clean training examples at train time, as this substantially increases the size of mention overlaps, leading to a simpler coreference resolution task.

5 Conclusions

Coreference resolution have the potential to help downstream NLP systems solve problems that require text understanding. However, the performance scores on the CoNLL benchmark are inflated, because mentions are largely overlapping in the whole dataset, and the evaluation in a constrained domain fails to expose the limitations of coreference resolvers in the wild. Our experiments

show that published resolvers fail to link coreferent mentions involving minor low-level noise and lexical changes. Beyond that, we show a caveat when mitigating the obstacles via adversarial approaches: lexical overlaps introduced by data augmentation must be removed from adversarial training and test sets so as to see how the approaches perform realistically.

Acknowledgments

The authors would like to thank Mark-Christoph Müller, Yufang Hou, Nafise Sadat Moosavi and the anonymous reviewers for their helpful comments and feedbacks. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. Haixia Chai has been supported by a Heidelberg Institute for Theoretical Studies PhD. scholarship. The contribution of Wei Zhao is supported by German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Granada.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *International conference on intelligent text processing and computational linguistics*, pages 136–145. Springer.

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. 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)*, pages 1405–1415.

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. Association for Computational Linguistics.

M Davis. 2003. Aoccdrnig to a rscheearch at cmabrigde uinervtisy. retrieved july 25, 2005.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Steffen Eger and Yannik Benz. 2020. From hero to zéroe: A benchmark of low-level adversarial attacks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. 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)*, pages 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. 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)*, pages 5802–5807, Hong Kong, China. Association for Computational Linguistics.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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, Vancouver, Canada. 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. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Sanjay Subramanian and Dan Roth. 2019. Improving generalization in coreference resolution via adversarial training. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 192–197, Minneapolis, Minnesota. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Dian Yu and Heng Ji. 2016. Unsupervised person slot filling based on graph mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44–53, Berlin, Germany. Association for Computational Linguistics.