Adapting Coreference Resolution to Twitter Conversations

Berfin Aktaş*
SFB1287, Research Focus Cognitive Sciences
University of Potsdam, Germany
berfinaktas@uni-potsdam.de

Veronika Solopova*
Free University of Berlin,
Germany
solopov97@zedat.fu-berlin.de

Annalena Kohnert
Department of Language Science and Technology
Saarland University, Germany
annalenakohnert@gmail.com

Manfred Stede
SFB1287, Research Focus Cognitive Sciences
University of Potsdam, Germany
stede@uni-potsdam.de

Abstract

The performance of standard coreference resolution is known to drop significantly on Twitter texts. We improve the performance of the (Lee et al., 2018) system, which is originally trained on OntoNotes, by retraining on manually-annotated Twitter conversation data. Further experiments by combining different portions of OntoNotes with Twitter data show that selecting text genres for the training data can beat the mere maximization of training data amount. In addition, we inspect several phenomena such as the role of deictic pronouns in conversational data, and present additional results for variant settings. Our best configuration improves the performance of the "out of the box" system by 21.6%.

1 Introduction and Related Work

Twitter messages present a discourse genre that includes noisy informal language with abbreviations and purposeful typos, use of nonstandard symbols such as # and @ signs, unintended misspellings, etc., which makes them challenging for NLP applications. We are here interested in the task of automated coreference resolution for nominal mentions in Twitter conversations, i.e., threads of messages that specifically reply to one another. In addition to non-standard words, Twitter conversations also show peculiar phenomena of referring, such as exophoric pointers to non-linguistic content in attached visual media, and mixed pronominal references to the same entity due to the nature of multi-user conversations (Aktaş et al., 2018).

Thus, tweets are a complicated genre for coreference resolution, but at the same time highly relevant for many applications that seek to extract information or opinions from users' messages. In this paper, we use a state-of-the-art resolution system built with the OntoNotes corpus (Pradhan et al., 2007) and experiment with adding annotated Twitter conversations to the training data. Next, we consider the different – spoken and written – genres included in the OntoNotes corpus. We thus conduct experiments with training on different portions, and we show that carefully selecting genre subsets beats the straightforward "taking as much as possible". Overall, our best configuration improves the "out of the box" performance of the system by Lee et al. (2018) on Twitter data by 21.6%.

To our knowledge, there is no work specifically on adapting coreference resolution to Twitter, other than the aforementioned study of Aktaş et al. (2018), which showed a significant drop in performance when a system with OntoNotes models is applied to Twitter. More generally, one of the few studies on domain adaptation for coreference resolution is (Do et al., 2015), which adapts the Berkeley system (Durrett and Klein, 2013) to narrative stories. Do et al. do not retrain the system but add linguistic features of narratives as soft constraints to the resolver. – At the same time, Twitter-adaptation has been investigated for other NLP tasks, such as NER. As an example, in (Ritter et al., 2011), performance is measured using tools trained with Twitter-related and out-of-domain data.

Regarding OntoNotes genre differences, Uryupina and Poesio (2012) and Pradhan et al. (2013) report varying performance in coreference resolution for distinct corpus sections; this work inspired our experiments reported in the following. Section 2 describes our data sets, and Section 3 the experiments. Section 4 provides various additional analyses that shed light on the domain adaptation problem, and Section 5 concludes.

* indicates equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2454–2460

2 Data

For our experiments,[1] we use the English portion of the OntoNotes benchmark used as training set in the CoNLL-2012 shared task (Pradhan et al., 2012). It has texts from spoken and written registers, and contains gold annotations at different layers, including coreference chains, i.e., sets of mentions referring to the same entity. Spoken data includes telephone conversations (**tc**), broadcast conversations (**bc**), and broadcast news (**bn**); written data contains magazine (**mz**), newswire (**nw**), pivot text (**pt**) and web blogs (**wb**). As shown in Table 1, the **ONT** corpus contains 1289K *tokens* in 2632 *documents* (in CoNLL terminology, documents are the units of independent annotation).

	docs	tokens	chains	mentions
ONT	2632	1289K	34K	152K
tc	111	81K	1931	12K
bc	284	144K	4236	18K
bn	711	172K	6138	21K
mz	410	164K	3534	13K
nw	622	387K	9404	34K
pt	320	210K	6611	42K
wb	174	131K	2993	12K
TW'	185	48K	1534	6K

Table 1: Corpus size and basic coreference statistics

Our second dataset is the Twitter Conversation corpus (**TW**) presented in (Aktaş et al., 2018). They are tree structures where each tweet has a parent (i.e. the tweet it is replied-to) except for the initial tweet starting the conversation. A tree can be shallow, with many replies on just one level, or it can be deep when participants interact with each other across several turns. The corpus holds 1756 tweets in 185 threads, defined as a path from the root to a leaf node of a conversation tree.[2] 69% of the coreference chains in this dataset contain coreferential relations across tweets. Hence, considering conversation context is important. We illustrate a thread structure with one example of coreference chain annotation in Figure 1.

The original TW corpus was annotated with a scheme slightly different from that of ONT. For systematic comparison, we modified the TW annotations so that they are conceptually parallel to

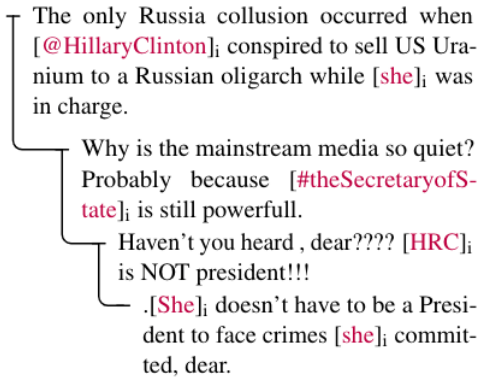

Figure 1: A thread sample in TW

ONT; we thus call the dataset **TW'** here.

3 Experiments

For our experiments, we chose 'e2e-coref' (Lee et al., 2018), an update of the end-to-end neural coreference resolver presented at EMNLP 2017. It introduced a refined approach based on differentiable approximation to higher-order inference, and ELMo embeddings (Peters et al., 2018) for span scoring, which significantly improved performance on English ONT. The approach achieved 73.0 F1, representing the 2018 state-of-art. Due to its cost efficiency, speed and flexibility, it was later used as basis for several recent state-of-art models, including SpanBERT (Joshi et al., 2020).

3.1 Test set

	Tokens	**Chains**	**Mentions**
train	44885	1411	5946
test	3260	123	408

Table 2: Twitter train/test distribution

Our main goal is to see how different training set configurations affect the coreference resolution performance on Twitter data. In order to achieve informative results, as the data is not linearly distributed and highly variable, we selected a representative test set not via random sampling, but through statistical analysis of three features: **number of tokens, chains** and **mentions** per document. To faithfully represent threads of all lengths, we determined the documents where these variables are situated either on the median, or in the first and fourth quartiles of the respective distribution, while omitting obvious outliers (see Figure 2). Because of the linear correlation of the three parameters

[1] Data distribution and scripts can be found at https://github.com/verosol/e2e-coref-to-Twitter

[2] Only the longest path has been used from each tree, so there is no redundancy in the data.

shown on Figure 3, we could make sure to only select the documents where all three are in the same range of their distributions.

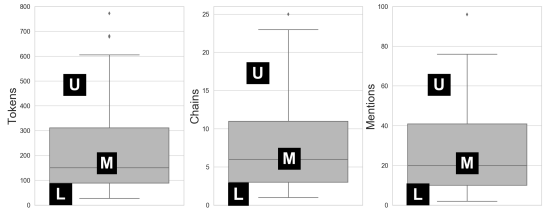

Figure 2: Distribution of the three considered parameters. U, L, M marks the forth (upper), first (lower) quartiles, and median respectively.

Among the pre-screened files, we checked each document, marking features of the annotated mentions (person, number, gender) as well as Twitter phenomena (hash-tags, user names, pronouns with typos, etc.). With this information, we excluded threads without enough coverage and variability of the phenomena in focus. As the threads are not evenly distributed in their total length, we compared the average, median and sum for each of the three characteristics in the whole corpus with those of the determined test set, confirming that all values lie under the 15% threshold of the total number. The final distribution is shown in Table 2.

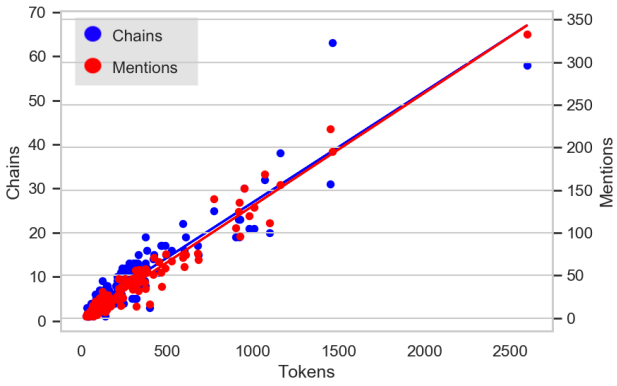

Figure 3: Each blue data point represents the chains and token count for each document, while red points denote mention and token information of the same documents.

3.2 Baseline Experiments

For evaluation, we use the official CoNLL-2012 scripts, measuring the average of precision, recall and F1 for muc, b3 and ceafe metrics. After we successfully reproduced the published e2e-coref results, we measured how a model trained on ONT performs on our Twitter test set (**Test A**). The resulting 45.18 F1 (see Table 4) is almost 28% lower than the result reported on the official ONT test set.

Test	Tokens	Chains	Mentions
A - ONT	1289K	34K	152K
B - TW only	44.8K	1.4K	5.9K
C - TW+ONT	1333.8K	35.4K	157.9K
D - TW+spok	269.8K	7.5K	35.9K
E - TW+writ	269K	5.8K	22.8K

Table 3: Experimental setup

A second baseline results from using only the TW' twitter corpus as training data, which lead to 60.8 F1 (**Test B**). Although this model is based on a rather small training set, it already improves significantly on baseline A and points to the difference between in-domain and out-domain training.

3.3 Effects of selecting training (sub-)sets

Noting that the presence of Twitter data in the training set is beneficial, for **Test C** we merged ONT and TW', with the latter forming 3.35% of the total size (see Table 3). The results show not only a performance increase of 17% in comparison to Test A, but also a 2% gain over Test B, demonstrating that combinations of both ONT and TW' can be crucial for the learning effects. To study this in more detail, we measured how performance on the test set reacts to training on different subsets of ONT. We roughly distinguished **spoken**, spontaneous language from **written** or edited texts.

Hence, in **Test D**, the training set consists of Twitter and only ONT's spoken genres, viz. broadcasts conversations and telephone conversations. As a consequence, the proportion of Twitter data in the training set rises from 3.35% to 16.6%. We found an increase in overall performance by 4.3%, indicating that the written genres may rather add confusion instead of benefit to this task. However, it is not entirely clear whether the improvement results from excluding the written genres or from increasing the proportion of Twitter data.

To answer this question, we proceeded to **Test E**, which combines the proportion of Twitter data present in Test D with documents from the written genres; we chose newswires (nw) and magazines (mz). Test E scores F1 61.25, which is 5.5% lower than Test D. This result may partly be due to the sparsity of the written data, with a smaller amount of chains and mentions present in the written genre

Test	Rec.	Prec.	F1	Rec.[1]	Prec.[1]	F1[1]	Rec.[2]	Prec.[2]	F1[2]
MUC									
A - ONT	38.24	55.89	45.41	35.74	51.36	42.15	41.05	66.47	50.75
B - TW only	56.84	74.65	64.54	50.95	70.89	59.29	-	-	-
C - TW+ONT	60.35	71.07	65.27	46.38	67.77	55.07	62.8	73.06	67.54
D - TW+spok	62.1	77.97	68.41	47.9	75.44	58.6	61.75	72.72	66.79
E - TW+writ	60.35	71.36	65.39	54.75	69.23	61.14	62.45	73.85	67.68
B^3									
A - ONT	35.14	56.02	43.18	33.19	51.68	40.42	37.21	66.78	47.79
B - TW only	51.64	68.77	58.99	46.31	63.52	53.57	-	-	-
C - TW+ONT	55.95	66.02	60.57	44.58	63.04	52.23	58.29	68.97	63.18
D - TW+spok	58.25	74.16	65.25	46.46	71.45	56.31	57.16	68.48	62.31
E - TW+writ	55.19	63.9	59.23	49.28	60.4	54.28	59.24	68.85	63.68
CEAFE									
A - ONT	44.5	49.76	46.98	43.26	47.59	45.32	49.13	61.04	54.44
B - TW only	50.97	69.66	58.87	44.54	65.96	52.96	-	-	-
C - TW+ONT	56.68	67.68	61.69	50.0	65.48	56.71	59.29	70.12	64.25
D - TW+spok	61.81	71.06	66.12	53.94	68.2	60.24	59.64	64.92	62.17
E - TW+writ	52.4	67.85	59.13	46.01	64.06	53.55	58.14	67.47	62.46
Average									
A - ONT	39.29	53.89	45.18	37.39	50.21	42.6	42.46	64.76	50.99
B - TW only	53.15	71.025	60.8	47.27	66.58	55.27	-	-	-
C - TW+ONT	57.76	68.25	62.51	46.9	65.43	54.67	60.12	70.71	**65.0**
D - TW+spok	**60.72**	**74.39**	**66.8**	49.43	**71.69**	**58.3**	59.51	68.7	63.76
E - TW+writ	55.98	67.7	61.25	**50.01**	64.56	56.32	59.94	70.05	64.60

Table 4: Results ($F1^1$, $F1^2$ are calculated after removing first and second person pronouns, and verb mentions respectively. They are discussed in Section 4)

documents (cf. Table 3), but still indicates an advantage of the spoken portion of ONT over the written one.

4 Additional Analyses

To gain further insight into the adaptation of coreference resolution to Twitter, we quantitatively and qualitatively compare the results of the best-performing test (D) to the baselines (see Table 5).

Mention length For all tests, the average token length of mentions additionally predicted by the system (spurious predictions) is significantly longer ($p \leq 0.05$) than that of the correct predictions. The higher the proportion of ONT training data (whose mentions are on avg. 0.72 tokens longer than in TW'), the longer those predictions are. At the same time they are significantly shorter ($p \leq 0.05$) than the missed gold predictions. Hence there is a tendency to select longer spans (especially when training on ONT), but these are also more error-prone.

Twitter-specific tokens Hashtags and usernames caused many errors in Test A. In tweets that are replies, user addresses are inserted at the beginning, so the majority of such tweet-initial usernames are not part of the syntax and have not been annotated. Table 5 shows that many of those names are incorrectly detected as mentions, while hashtags are completely ignored. With Twitter training data in Test B, identification of Twitter-specific tokens works better. Tweet-initial usernames are ignored as mentions and some username and hashtags are now correctly predicted. Test D shows further improvements for syntactically-integrated hashtags, but usernames or non-integrated hashtags still remain unresolved.

Pronouns Although they are relatively evenly distributed in the gold annotations, more 3rd person pronouns are resolved than 1st and 2nd ps. pronouns in Test A, resulting in an overall F1 of 0.769. In Test B with Twitter training data, which is rich in pronouns, pronoun performance improves for 1st and especially 2nd ps., and remains the same for 3rd ps., improving the F1 to 0.917. In Test D, pronoun performance is slightly worse (0.905).

As the entire training data in B and D is conversational, which by nature has many 1st and 2nd ps. pronouns, we repeated all test with removing those chains containing only 1st and 2nd ps. pronouns. This is to make sure that improvement is not exclusively caused by easy detection of the pronouns. The results are in column F1[1] in Table 4. While deictic pronouns have a major impact on F1, we still see improvements over the baseline for all tests but C, meaning that generally, detection of other anaphoric expressions improves as well.

Verb annotations Verb mentions are possible in ONT if they co-refer with a nominal mention (Pradhan et al., 2007), but they are not annotated in TW'. Thus four predicted verb mentions in Test A, of which two are correctly linked with the demonstrative pronoun *that*, are counted as erroneous predictions. After adding training data from TW' in Test D however, no verbal mentions are predicted. To check the influence of this annotation difference, we also ran all tests with the verbal annotations removed from ONT, which reduced mentions by 2.4% and chains by 3.6%. Column F1[2] in Table 4 shows the results. While training with only spoken genres outperformed more written dominant training data in previous experiments, we now see the opposite with Test D giving the worst results. These variations motivate looking further into the specific effects of different training data combinations and how verb annotations (both generally and depending on text genres) influence an otherwise purely nominal coreference resolution task.

Chain Linking The last section of Table 5 shows that Test B improves the number of correctly predicted chains compared to Test A, and it further increases in Test D, almost doubling from Test A. Partially correct chains also increase over the tests, and the number of missed entities (cases where not a single mention of an entity is predicted) is reduced by 51.3%. Notably, chains consisting only of identical strings profited the most from the combined training set in D.

	Gold	A	B	D
Pred. Mentions	408	305	307	334
Usernames	8	51	6	5
tweet-initial	1	44	0	0
Hashtags	11	0	4	5
Correctly Pred.	408	218	265	293
Avg. #tokens	1.64	1.41	1.13	1.18
Pronouns	219	149	199	194
1st person	57	38	53	50
2nd person	64	26	63	62
3rd person	68	60	61	59
Usernames	8	6	5	5
tweet-initial	1	1	0	0
Hashtags	11	0	3	5
Pred. Chains	123	110	90	107
Correct Chains	-	18	27	37
Partially Correct[3]	-	10	11	14
Missed Entities	-	39	32	20

Table 5: Properties of predicted mentions and chains

5 Conclusion

We showed that the performance of a state-of-the-art "standard" coreference resolution system run on Twitter conversations can improve by 21.6% by adding in-domain training data. In fact, even small amounts of added in-domain data can have an impact. Further, interestingly, for the out-domain training data (ONT), the choice of genre can make a bigger difference than the bare amount of data. Our additional analyses considered two more variants of the main experiment design: While all results given in Table 4 indicate that adding Twitter data to the training set improves the performance significantly, the best combination of in-domain and out-domain data can depend on specific factors as discussed in section 4. Also, we showed that improvements from Twitter training data do not result just from the large proportion of 1st and 2nd ps. pronouns (as one might have wondered). Finally, we tested the effect of removing verb mentions from ONT, which exhibits different patterns than other setups regarding the best combination of training data. The result encourages deeper exploration of training data arrangements in terms of these features.

In future work we plan to focus more on the specific kinds of training data portions and examine the influence of spoken versus written register, and on that of formal versus informal language (which need not necessarily coincide).

[3] All gold mentions found, but also spurious mentions.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 317633480 - SFB 1287, Project A03.

References

Berfin Aktaş, Tatjana Scheffler, and Manfred Stede. 2018. Anaphora Resolution for Twitter Conversations: An Exploratory Study. In *Proceedings of the Workshop on Computational Models of Reference, Anaphora, and Coreference, CRAC@HLT-NAACL 2018*, New Orleans, Louisiana. Association for Computational Linguistics.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. 2015. Adapting Coreference Resolution for Narrative Processing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2267, Lisbon, Portugal. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2013. Easy Victories and Uphill Battles in Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards Robust Linguistic Analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152. 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. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica Macbride, and Linnea Micciulla. 2007. Unrestricted Coreference: Identifying Entities and Events in OntoNotes. *International Conference on Semantic Computing*, 0:446–453.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Olga Uryupina and Massimo Poesio. 2012. Domain-specific vs. Uniform Modeling for Coreference Resolution. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 187–191, Istanbul, Turkey. European Language Resources Association (ELRA).

A Appendix: Alignment of Annotation Schemes

We use two corpora, Twitter Conversations (TW) and OntoNotes, in the experiments presented in the paper. Only the *identity* relations are annotated in both of the corpora and mentions building singleton chains (i.e. chains containing only 1 item) are not considered as markables in either of them. However their annotation schemes are not fully aligned; there exist differences in the definition of markables. For the sake of comparability of the experimental results, we aligned the type of annotated markables as much as possible by applying semi-automated procedures. We summarize below the main differences we determined and applied handling strategies to harmonize them:

- In TW, predicative nouns (e.g. *This is* [*a fake account*]), and headless relative clauses having the grammatical role of a noun phrase (e.g. *A mature male kangaroo doing* [*what*] *it's built for*) are considered as markables, but not so in OntoNotes. We removed the predicative noun and relative pronoun annotations in TW.

- In TW, appositions (e.g. [*His wife*], [*Florence*], *fell ill.*) are annotated separate from the preceding noun they co-refer with. In the CoNLL formatted version of OntoNotes that we use, appositions are merged with the nominals they modify (e.g. e.g. [*His wife, Florence*], *fell ill.*). Therefore, the appositive modifiers in TW are merged with the preceding co-referring noun phrase.

- Generic ”you” instances are annotated in TW but not in OntoNotes. We removed generic ”you” annotations from TW.

- In TW, ”reflexives” are annotated as separate mentions even if they are used for focus (e.g. [*The president*] [*himself*] *said this*). However, the focus reflexives are both annotated as a separate markable and also a part of the span of the preceding co-referring noun phrase in OntoNotes (e.g. [*The president* [*himself*]] *said this*). Therefore, the focus reflexives in TW are added to the span of the preceding co-referring noun phrase.

If the removal of a mention made the remaining chain a singleton (i.e. only 1 mention left in the chain), the whole chain is removed from the annotations, as no singleton chains are allowed in the OntoNotes scheme.

B Appendix: Preprocessing the Data

In TW dataset:

- We normalized parentheses, namely left and right bracket tokens into ’-LRB-’ and ’-RRB-’, respectively.

- We converted all smiley and emoji tokens into the strings of ”%smiley” and ”%emoji”, respectively.

- We did not apply any preprocessing to hashtags and @-usernames.

C Appendix: Experimental Setup

The experiments are conducted on two servers with GPU, GeForce GTX 1080.