Investigating Domain-Specific Information for Neural Coreference Resolution on Biomedical Texts

Hai-Long Trieu[1], **Nhung T. H. Nguyen**[2], **Makoto Miwa**[1,3] and **Sophia Ananiadou**[2]
[1]Artificial Intelligence Research Center (AIRC),
National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2]National Centre for Text Mining, University of Manchester, United Kingdom
[3]Toyota Technological Institute, Japan
long.trieu@aist.go.jp, makoto-miwa@toyota-ti.ac.jp
{nhung.nguyen, Sophia.Ananiadou}@manchester.ac.uk

Abstract

Existing biomedical coreference resolution systems depend on features and/or rules based on syntactic parsers. In this paper, we investigate the utility of the state-of-the-art general domain neural coreference resolution system on biomedical texts. The system is an end-to-end system without depending on any syntactic parsers. We also investigate the domain specific features to enhance the system for biomedical texts. Experimental results on the BioNLP Protein Coreference dataset and the CRAFT corpus show that, with no parser information, the adapted system compared favorably with the systems that depend on parser information on these datasets, achieving 51.23% on the BioNLP dataset and 36.33% on the CRAFT corpus in F1 score. In-domain embeddings and domain-specific features helped improve the performance on the BioNLP dataset, but they did not on the CRAFT corpus.

1 Introduction

Deep neural systems have recently achieved the state-of-the-art performance on coreference resolution tasks in the general domain (Clark and Manning, 2016; Wiseman et al., 2016; Lee et al., 2017). These systems do not heavily rely on manual features since the networks automatically build advanced features from the input. Such an attribute has made deep neural systems preferable to traditional manual feature-based systems.

In the biomedical domain, coreference information has been shown to enhance the performance of entity and event extraction (Miwa et al., 2012; Choi et al., 2016a). Most of work in this domain use rule-based or hybrid approaches (Nguyen et al., 2011, 2012; Miwa et al., 2012; D'Souza and Ng, 2012; Li et al., 2014; Choi et al., 2016b; Cohen et al., 2017). These systems rely on syntactic parsers to extract hand-crafted features and rules, e.g., rules based on predicate argument structure (Nguyen et al., 2012; Miwa et al., 2012) or features based on syntax trees (D'Souza and Ng, 2012). These rules are designed specifically for each type of coreference, such as noun phrases, relative pronouns, and non-relative pronouns. Moreover, several rules are restricted to specific entities of the training corpus, e.g., protein entities for the BioNLP Protein Coreference dataset (Nguyen et al., 2011).[1]

Given the fact that deep learning methods can produce the state-of-the-art performance on general texts, we are motivated to apply such methods to biomedical texts. We therefore raise three research questions in this paper:

- How does a general domain neural system with no parser information perform on biomedical domain?
- How we can incorporate domain-specific information into the neural system?
- Which performance range the system is in comparison with existing systems?

In order to address these questions, we directly apply the end-to-end neural coreference resolution system by Lee et al. (2017) (Lee2017) to biomedical texts. We then investigate domain specific features such as domain-specific word embeddings, grammatical number agreements between mentions, i.e., mentions are singular or plural, and agreements of MetaMap (Aronson and Lang, 2010) entity tags of mentions. These features do not rely on any syntactic parsers. Moreover, these features are also general for any biomedical corpora and not restricted to the corpora we use.

[1]http://2011.bionlp-st.org/home/protein-gene-coreference-task

Proceedings of the BioNLP 2018 workshop, pages 183–188
Melbourne, Australia, July 19, 2018.

We evaluated the Lee2017 system on two datasets: the BioNLP Protein Coreference dataset (Nguyen et al., 2011) and CRAFT (Cohen et al., 2017). Our experimental results have revealed that the system could achieve reasonable performance on both corpora. The system outperformed several systems on the BioNLP dataset that employed rule-based (Choi et al., 2016b) and conventional machine learning methods (Nguyen et al., 2011) using parser information, although it was not competitive with the state-of-the-art systems. Integrating in-domain embeddings and domain-specific features into the deep neural system improved the performance of both mention detection and mention linking on the BioNLP dataset, but the integration could not enhance the performance on the CRAFT corpus.

2 Methods

In this section, we briefly introduce the baseline Lee2017 system (Lee et al., 2017) and present domain-specific features to adapt the system to biomedical texts.

2.1 Baseline System

The baseline Lee2017 system treats all spans up to the maximum length as mention candidates. Each mention candidate is represented as a concatenated vector of the first word, the last word, the soft head word, and the span length embeddings. The embeddings for the first and last words are calculated from the outputs of LSTMs (Hochreiter and Schmidhuber, 1997), while those for soft head word are calculated from the weighted sum of the embeddings of words in the span using an attention mechanism (Bahdanau et al., 2014). These candidates are ranked based on their mention scores s_m calculated as follows:

$$\mathrm{s}_m(i) = w_m \cdot \mathrm{FFNN}_m(g_i), \quad (1)$$

where w_m is a weight vector, FFNN denotes a feed-forward neural network, and g_i is the vector representation of a mention i.

After mentions are decided, the system resolves coreference by linking mentions back to their antecedent using antecedent scores s_a calculated as:

$$\mathrm{s}_a(i,j) = w_a \cdot \mathrm{FFNN}_a([g_i, g_j, g_i \circ g_j, \phi(i,j)]), \quad (2)$$

where $\circ$ denotes an element-wise multiplication and $\phi(i,j)$ represents the feature vector between the two mentions.

2.2 Domain-specific features

We incorporate the following domain-specific features to enhance the baseline system.

In-domain word embeddings: The input word embeddings play an important role in deep learning. Instead of using embeddings trained on general domains, e.g., word embeddings provided with the word2vec tool (Mikolov et al., 2013), we use 200-dimensional embeddings trained on the whole PubMed and PubMed Central Open Access subset (PMC) with a window size of 2 (Chiu et al., 2016).

Grammatical numbers: We check mentions' grammatical numbers, i.e., whether each mention is singular or plural. A mention is singular if its part-of-speech tag is NN or if it is one of the five singular pronouns: *it*, *its*, *itself*, *this*, and *that*. A mention is plural if its part-of-speech tag is NNS or if it is one of the seven plural pronouns: *they*, *their*, *theirs*, *them*, *themselves*, *these*, and *those*.

MetaMap entity tags: We employ MetaMapLite[2] to identify all possible entities according to the UMLS semantic types.[3] In cases that MetaMapLite assigns multiple semantic types for each entity, we take into account all of the types.

The grammatical numbers and MetaMap entity tags are incorporated into the network as follows. We firstly pre-processed the input and assigned token-based values for each type of features. For example, a token may have "singular", "plural", or "unknown" as the number attribute. Meanwhile, the MetaMap entity tags are distributed to each token with their position information chosen from "Begin" and "Inside". These features are finally encoded as a binary vector of $\phi(i,j)$ in Equation 2 that shows whether two mentions i and j has the number agreement and whether they share the same MetaMap semantic type.

3 Experiments

3.1 Data

We employed two biomedical corpora: BioNLP Protein Coreference dataset (Nguyen et al., 2011) and CRAFT (Cohen et al., 2017). The BioNLP dataset consists of 1,210 PubMed abstracts selected from the GENIA-MedCo coreference corpus. CRAFT (Cohen et al., 2017) provides coref-

[2] https://metamap.nlm.nih.gov/MetaMapLite.shtml

[3] https://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt

	BioNLP	CRAFT
Training set (docs)	800	54
Development set (docs)	150	6
Test set (docs)	260	7
Avg. sent. per doc	9.15	274.75
Avg. words per doc	258.00	8,060.85
Vocabulary size	15,900	27,405

Table 1: Characteristics of BioNLP and CRAFT.

erence annotations of 67 full papers extracted from PMC. While BioNLP focusses on protein/gene coreference, CRAFT covers a wider range of coreference relations such as events, pronominal anaphora, noun phrases, verbs, and nominal premodifiers corefernce. In the CRAFT corpus, coreference is divided into two types: identity chains (a set of base noun phrases and/or appositives that refer to the same thing in the world) and appositive relations (two noun phrases that are adjacent and not linked by a copula). We use only the identity chains.

The BioNLP dataset was officially divided into training, development, and test sets. Regarding CRAFT, we randomly divided it into three subsets in a ratio of 8:1:1 for training, development, and test, respectively. Detailed characteristics of the two corpora as well as these three sets are reported in Table 1. It is noticeable that CRAFT is a corpus of full papers, which makes it more challenging for text mining tools than the BioNLP dataset—a corpus of abstracts (Cohen et al., 2010).

3.2 Settings

We first directly applied the Lee2017 system to the corpora. Lee2017 used two pretrained embeddings in general domains provided by Pennington et al. (2014) and Turian et al. (2010), and all default features such as speaker, genre, and distance.

To train the Lee2017 system, we employed the same hyper-parameters as reported in Lee et al. (2017) except for a threshold ratio. Although Lee2017 used the ratio $\lambda = 0.4$ to reduce the number of mentions from the list of candidates, we tuned it on the BioNLP development set and used $\lambda = 0.7$.

We then investigate the impact of each feature on the biomedical texts by preparing the following four systems:

- Lee2017: general embeddings, speaker, genre, and distance features
- PubMed: biomedical embeddings, same features as Lee2017
- PubMed-SG: PubMed with no speaker and genre features
- PubMed+*: PubMed with the MetaMap feature (MM) and/or the grammatical number feature (Num).

BioNLP	Prec.	Rec.	F1 (%)
Lee2017	81.15	63.81	71.44
PubMed	81.01	66.12	72.81
PubMed-SG	79.23	65.73	71.85
PubMed+MM	80.41	**67.17**	73.20
PubMed+Num	**81.91**	66.31	**73.29**
PubMed+MM+Num	81.04	66.69	73.17
CRAFT	**Prec.**	**Rec.**	**F1 (%)**
Lee2017	70.76	48.71	57.70
PubMed	70.93	46.90	56.46
PubMed-SG	71.98	**50.24**	**59.18**
PubMed+MM	71.11	47.91	57.25
PubMed+Num	**72.79**	42.55	53.70
PubMed+MM+Num	71.60	45.00	55.27

Table 2: Results of mention detection on the development set of BioNLP and CRAFT. The highest numbers are shown in bold.

For evaluation, we calculated precision, recall, and F1 on MUC, B^3, and $CEAF_{\phi 4}$ using the CoNLL scorer (Pradhan et al., 2014). For the BioNLP dataset, we also employed the scorer provided by the shared task organisers to make fair comparisons with previous work. We reported the performance on two sub-tasks: (1) mention detection, i.e., to identify coreferent mentions, such as named entities, prepositions or noun phrases, and (2) mention linking, i.e., to link these mentions if they refer to the same thing. The result of the first task affects that of the second one.

3.3 Results

Results on the development sets of the two corpora are presented in Table 2 for mention detection and Table 3 for mention linking (see Appendix A for detailed scores in different metrics).

Regarding the BioNLP dataset, the Lee2017 system performed reasonably well even when it did not use any domain-specific features. Replacing general embeddings by the biomedical ones improved F1 score in general (Lee2017 v.s. PubMed). Removing speaker and genre features (-SG) did not help enhance the performance.

System	BioNLP	CRAFT
Lee2017	61.25	33.85
PubMed	62.51	33.92
PubMed-SG	61.47	**34.85**
PubMed+MM	**63.41**	33.91
PubMed+Num	63.16	31.28
PubMed+MM+Num	63.12	32.77

Table 3: Average F1 scores (%) of mention linking on the development set of BioNLP and CRAFT.

Adding MetaMap's tags (+MM) or the number feature (+Num) produced slightly better scores in comparison to PubMed. However, combining the two features at the same time was not as effective as expected. Among the proposed features, the agreement on MetaMap entity tags (+MM) was the strongest one on the BioNLP dataset.

The impact of the features was quite different on the CRAFT corpus. As shown in Table 2, introducing biomedical embeddings (PubMed) show slightly worse F1 score on mention detection than Lee2017 but it also show a slight improvement on mention linking. Removing speaker and genre features (-SG) boosted the performance. However, adding domain-specific features all harmed the performance. As a result, PubMed-SG showed the best score on the CRAFT development set.

Results in Tables 2 and 3 justify the fact that the CRAFT corpus is more challenging than the BioNLP dataset. The scores of the experimented systems on the CRAFT corpus were always lower than those on the BioNLP dataset. This is reasonable because (1) CRAFT consists of full papers that are significantly longer than abstracts, (2) it covers a wide range of anaphors, and (3) its identity chains can be arbitrarily long.

We applied the best performing system on each development set, i.e., PubMed+MM for BioNLP and PubMed-SG for CRAFT, to its test set, and reported the results in Tables 4 and 5 with showing the performance in previous work for comparison. Table 4 reveals that the neural system outperformed five systems that used SVM and rule-based approaches including the best system on the shared task, and the system could compete with Nguyen et al. (2012)'s. Meanwhile, on the CRAFT corpus (Table 5), we could only produce better performance than the general state-of-the-art system, especially due to the low precision.

System	Prec	Rec	F1 (%)
TEES (BioNLP ST)	67.2	14.4	23.8
ConcordU (BioNLP ST)	63.2	19.4	29.7
UZurich (BioNLP ST)	55.5	21.5	31.0
UUtah (BioNLP ST)	73.3	22.2	34.1
Choi et al. (2016b)	46.3	50.0	48.1
PubMed+MM	55.6	47.5	51.2
Nguyen et al. (2012)	50.2	52.5	51.3
Miwa et al. (2012)	62.7	50.4	55.9
D'Souza and Ng (2012)	55.6	67.2	60.9

Table 4: Results of mention linking on the test set of the BioNLP dataset. The F-scores are in ascending order.

System	Prec.	Rec.	F1
General state-of-the-art	0.93	0.08	0.14
Rule-based	0.78	0.29	0.42
Union of the two output	0.78	0.35	0.46
PubMed-SG	0.44	0.31	0.36

Table 5: B^3 scores of mention linking on the CRAFT test set in comparison with the three systems by Cohen et al. (2017). This is not a fair comparison as our system only addressed identity chains and the test set is different from theirs.

4 Conclusion

We have applied a neural coreference system to biomedical texts and incorporated domain-specific features to enhance the performance. Experimental results on two biomedical corpora, the BioNLP dataset and the CRAFT corpus, have shown that (1) the neural system performed reasonably well with no parser information, (2) the in-domain embeddings and domain-specific features did not consistently perform well on the two corpora, and (3) the system could attain better performance than several rule-based and traditional machine learning-based systems on the BioNLP dataset.

As future work, we would like to investigate feature representations to make input features useful to a target domain. We will also incorporate rules in the existing systems into the network.

Acknowledgments

This research has been carried out with funding from AIRC/AIST and results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Billy Chiu, Gamal K. O. Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to Train good Word Embeddings for Biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174.

Miji Choi, Haibin Liu, William Baumgartner, Justin Zobel, and Karin Verspoor. 2016a. Coreference resolution improves extraction of biological expression language statements from texts. *Database*, 2016:baw076.

Miji Choi, Justin Zobel, and Karin Verspoor. 2016b. A categorical analysis of coreference resolution errors in biomedical texts. *Journal of biomedical informatics*, 60:309318.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. *CoRR*, abs/1606.01323.

K. Bretonnel Cohen, Helen L. Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E. Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11(1):492.

K. Bretonnel Cohen, Arrick Lanfranchi, Miji Jooyoung Choi, Michael Bada, William A. Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E. Hunter. 2017. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC Bioinformatics*, 18(1):372.

Jennifer D'Souza and Vincent Ng. 2012. Anaphora resolution in biomedical literature: a hybrid approach. In *BCB*, pages 113–122. ACM.

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

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. Association for Computational Linguistics.

Lishuang Li, Liuke Jin, Zhenchao Jiang, Jing Zhang, and Degen Huang. 2014. Coreference resolution in biomedical texts. In *BIBM*, pages 12–14. IEEE Computer Society.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–1765.

N. L. T. Nguyen, J.-D. Kim, and J. Tsujii. 2011. Overview of bionlp 2011 protein coreference shared task. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 74–82, Portland, Oregon, USA. Association for Computational Linguistics.

Ngan Nguyen, Jin-Dong Kim, Makoto Miwa, Takuya Matsuzaki, and Junichi Tsujii. 2012. Improving protein coreference resolution by simple semantic classification. *BMC Bioinformatics*, 13(1):304.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–1543.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word Representations: A Simple and General Method for Semi-supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. *CoRR*, abs/1604.03035.

A Detailed results

We report detailed results of mention linking on the development set of the two corpora in Table 6 and Table 7. Due to the long running time of the scorer, we were not able to report $\text{CEAF}_{\phi 4}$ scores for CRAFT.

System	MUC			B^3			$CEAF_{\phi 4}$			Avg.
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	F1 (%)
Baseline	65.31	45.03	53.30	71.50	50.27	59.03	77.06	66.54	71.41	61.25
PubMed	65.44	47.04	54.74	71.12	51.85	59.98	77.25	68.85	72.81	62.51
PubMed-SG	63.02	46.58	53.57	68.72	51.72	59.02	76.11	68.01	71.83	61.47
PubMed+MM	66.17	48.30	55.84	71.62	52.95	60.89	76.70	70.53	73.49	63.41
PubMed+Num	66.81	47.83	55.75	72.27	52.23	60.64	78.15	68.63	73.08	63.16
PubMed+MM+Num	65.73	47.37	55.06	71.68	52.66	60.72	77.53	70.04	73.59	63.12

Table 6: Results of mention linking on the BioNLP development set.

System	MUC			B^3			Avg. F1 (%)
	Prec.	Rec.	F1	Prec.	Rec.	F1	
Baseline	45.46	27.17	34.02	44.29	27.17	33.68	33.85
PubMed	47.36	27.34	34.67	44.89	26.30	33.17	33.92
PubMed-SG	46.04	28.49	35.20	43.33	28.67	34.50	34.85
PubMed+MM	46.22	27.37	34.38	43.70	27.09	33.44	33.91
PubMed+Num	47.31	23.40	31.31	48.70	23.01	31.25	31.28
PubMed+MM+Num	46.85	25.41	32.95	45.78	25.30	32.59	32.77

Table 7: Results of mention linking on the CRAFT development set.