A Deep Learning Framework for Coreference Resolution Based on Convolutional Neural Network

Jheng-Long Wu
Institute of Information Science
Academia Sinica
Taipei, Taiwan
E-mail: jlwu.studio@gmail.com

Wei-Yun Ma
Institute of Information Science
Academia Sinica
Taipei, Taiwan
E-mail: ma@iis.sinica.edu.tw

***Abstract*—Recently many researches have shown that word embeddings are able to represent information from word related contexts or its nearest neighborhood words, and thus are applied in many NLP tasks successfully. In this paper, we propose convolutional neural network model to extent word embeddings to mention/antecedent representation. These representations are obtained through convoluting neighboring word embeddings and other contextual information for coreference resolution. We evaluate our system on the English portion of the CoNLL 2012 Shared Task dataset and show that the proposed system achieves a competitive performance compared with the state-of-the-art approaches. We also show that our proposed model especially improves the coreference resolution of long spans significantly.**

Keywords- deep learning; word embeddings; mention embeddings; coreference resolution; convolutional neural network

I. INTRODUCTION

Coreference resolution is an important and challenging task in natural language processing (NLP). Entities and mentions of context often refer to the same entity in a document. The accurate detection of their coreference is difficult because the coreferent relationships are usually very dependent on various semantic, morph-syntax and gramma [1]-[2]. In recent years, many machine learning approaches rely on complex models to solve coreference resolution problems. For instance, mention-pair models aim to classify coreference relation of two mentions at one time. It is efficient and easy to implement, and also faster to train in contrast of other models. It relies on a binary classifier with some features of mentions to decide whether a mention and its antecedent entity were coreferent or not [3]-[6].

However, a challenge for mention-pair models is how to consider the overall semantics of mentions, i.e., the overall meaning of a mention which is composed of a sequence of words and mention pairs' context semantic information. Therefore, to consider the overall semantics of mentions, we present a convolutional neural networks (CNNs) model in deep learning framework to learn mentions' representation and coreference classification which are based on word embeddings. Our proposed framework can well consider and integrate string and numeric information. The coreference classification of our proposed model has the advantage of learning mention pair string representation (local information) through mention/antecedent word embeddings.

We believe our proposed model for textual data make three contributions to coreference resolution. First, through CNN, we can integrate the individual words' semantics into mentions' semantics. Second, we successfully applied the string information such as embeddings of dependency head words of mentions. Third, our experimental results show that each component of our deep neural network plays its role and benefit the whole performance of coreference resolution.

II. RELATED WORKS

For NLP tasks, a deep learning model usually involves the application of distributed word representations (word embeddings) obtained by an unsupervised learning method in neural network frameworks [7]-[8]. Word embedding is a vector representation that encodes semantic and syntactic features of words. Among various deep learning models, CNN features its convolving filters that are able to extract local features from nearest neighbor information and have been shown effective and promising results for many NLP tasks [9]-[10].

For coreference resolution problem, the rule-based approaches have stable performance in many cases but they have a major drawback that the rules are hand-crafted. Many slight relationships between mentions are not easily identified and maintained via human processing. Therefore, recent basic coreference approaches is mention pair model which is easy to implement and allows for fast prototyping [11]-[14]. More advanced approaches are using mention ranking model, which is based on mention-pair model but only select the highest coreferent score among all forward antecedents of a mention. In recent years, the neural network based coreference resolution systems have been proposed, where the mention pair models or mention ranking models are usually used to design to solve the coreference resolution problems i.e., [5]-[6] and [15]-[18]. Of the mention-pair framework, singleton mention identification is usually a key process for mention detection [1]. An effective singleton classifier can discard singleton entities before coreference resolving. Our proposed deep learning framework also consists of a singleton classification stage. For coreference classifier, we learn the semantic information of a mention as well as its dependency word and use mention ranking model to determine which mention pair has the highest confidently coreference. We provide the implementation of our system at https://github.com/jlwustudio.

III. A DEEP LEARNING FRAMEWORK FOR COREFERENCE RESOLUTION

Our deep learning framework for coreference resolution is composed of multiple neural network architectures with a mention ranking mechanism. The system flow as show in Fig. 1 and includes three processes such as singleton classification, coreference classification and coreferent pair linking.

Figure 1. The system flow of the proposed coreference resolution system.

A. Singleton Classifier

The singleton classifier is shown in Fig. 2. This classifier considers five representations which are word, dependency, string, numeric, and mention. The word representation is constructed by CNNs with the input of the word embeddings of mention words. The construction process is similar with [9] in their sentence representation. The dependency, string, numeric, and mention representations are constructed by FCN. The inputs of the string representation are the concatenation of the dependency and word representations. The inputs of mention representation are the concatenation of the string and numeric representations. If mentions are not singleton from singleton classification, then they further go into the next coreference classifier.

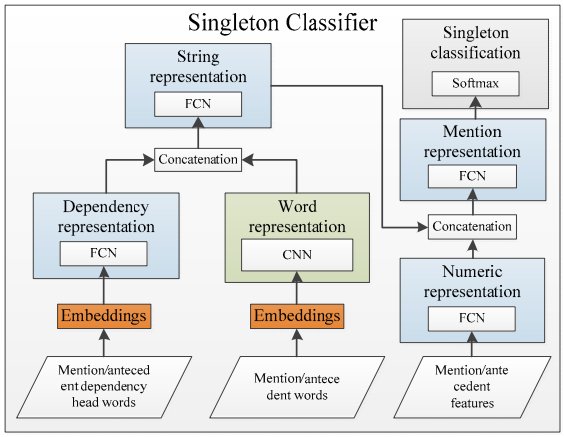

Figure 2. The processes of the proposed singleton classifier.

B. Coreference Classifier

The coreference classifier is shown in Fig. 3. This classifier considers six representations, which are dependency, antecedent, mention, pair string, pair numeric, and mention pair. The mention/antecedent representations are constructed by CNN. Other representations are constructed by FCN. The inputs of the mention pair string representation are the concatenation of the dependency, antecedent and mention representations. The input of the pair numeric representation is the concatenation of the antecedent mention pair features. The inputs of the mention pair representation are the concatenation of the pair string and pair numeric representations. The last process is coreference classification, which is a softmax function to provide the coreferent pair score.

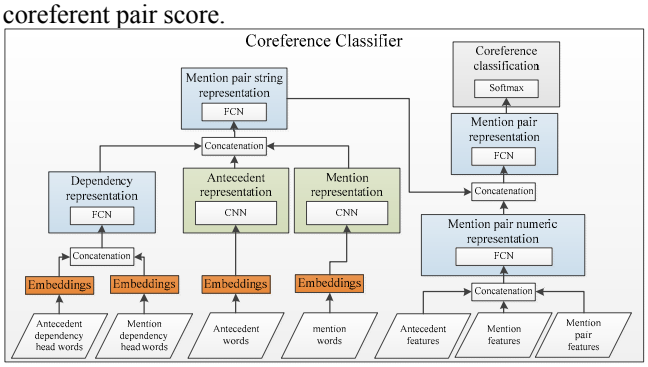

Figure 3. The processes of the proposed coreference classifier.

C. Coreferent Pair Linking (Mention Ranking Model)

The coreference classifier has predicted coreferent scores for all mention pairs of a targeted mention. We further find out which mention pair has the maximum score for the targeted mention. We apply mention ranking approach as a best-first to cluster mention and antecedent. For example, a mention m has a set of forward candidate antecedent $A(m)$, we refer to this set as m's potential antecedents. A mention-ranking model defines a score function $s(a,m)$ for a mention pair – (a,m), and the antecedent of highest score k^* for a mention m is found by,

$$k^* = \arg\max_{k \in A(m)} s(k,m) \quad \text{if } s(k,m) > \xi, \tag{1}$$

where the ξ is a coreferent threshold.

IV. EXPERIMENTS

In order to evaluate the performance of our proposed deep framework for coreference resolution, the CoNLL 2012 Share Task dataset [19] is used. There are 2,802 and 384 documents in training and testing, respectively. We extracted 0.95M coreferent mention pairs and 14M non-coreferent mention pairs from the training. To evaluate our system, three metrics - MUC [20], B-Cubed [21] and CEAF [22] are used. In addition, the optimization for optimizing weight W is Adam optimizer by [23].

A. Experimental Setup

There have 10 hyper-parameters are used in the paper. Dim. of layer on FCN and CNN is 200. Number of layer on FCN and CNN is 5. Filter size of CNN is 2. Batch size is 128. Embedding size is 300. We design the three different input feature sets for evaluate feature performances. The detailed of different input feature sets are as follows:

- **NF**: Numeric Features. 17 variables are used for a mention and the antecedent, and 5 relation variables are for a mention pair of a mention and its antecedent mention.

- **SF**: String Features. We use the word list and dependency words of mention. The word list is used for CNN to learn words representation. The dependency word is used for FCN to learn dependency representation.
- **SNF**: String with Numeric Features. We considered two features at the same time.
- **MENF**: Numeric Features with mean embedding of the mention words. Thus, the mention/antecedent representations without CNNs process.

The detailed features as follow: String feature have two features which are word list and dependency word. Numeric feature on mention have 17 features which are animate, entity, gender, head word index, NER, number of modifiers, number of predicates, number of word, person, pleonastic, POS of first word, POS of head word, position of head word in end of sentence, position of head word in first of sentence, position of head word in middle of sentence, position of head word in third of sentence and quantifier. Numeric feature on relation have 5 features which are distance of sentence, distance of word, same first word, same head word, and same speaker.

The two neural network structures of CNNs and FCNs have used to our proposed coreference resolution system including singleton and coreference. All initial word embeddings of a mention/singleton and their dependency words are obtained though looking up pre-trained word embeddings with 300d dimension. It trained from Wikipedia 2014 + Gigaword 5 corpus by Glove toolkit.

In singleton classifier, 5 layers CNNs are used to learn mention/singleton representation. 5 layers FCNs are used to converge the parameters of all string representation and other 5 layers FCNs are used to converge the parameters of numeric representation. Then we feed string and numeric representations into a 10 layers FCN to converge the parameters. The top layer is a softmax function, aiming to output the singleton probability. All mention/antecedent with singleton probability that are larger than 0.9 are considered to be non-singleton mention. In coreference classifier. 5 layers CNNs are used to learn mention embeddings. 5 layers FCNs are used to converge the parameters of all string representation and other 5 layers FCNs are used to converge the parameters of numeric representation. Then we feed two string and numeric representations into a 10 layers FCN to converge the parameters. The top layer is a softmax function, aiming to output the coreferent score. All mention pairs with coreferent scores that are larger than coreferent threshold 0.9 are considered to be coreferent pairs.

B. Results on Gold Mention

Table I shows the performance of different input feature sets. The results show that SNF has higher CoNLL score - 75.93 compared with the SF, NF and MENF. MENF has 74.06 of CoNLL score, which is also better than NF and SF. Comparing to MENF, our proposed SNF outperforms MEFN, which proves that CNNs indeed bring a certain level of benefits.

TABLE I. RESULT ON TEST SET WITH GOLD MENTION

Feature	MUC			B^3			CEAF			CoNLL
	P	R	F1	P	R	F1	P	R	F1	Mean
SF	80.69	57.72	67.30	75.38	56.60	64.65	31.76	62.15	42.04	57.99
NF	89.10	78.39	83.40	79.40	70.01	74.41	53.33	74.87	62.29	73.36
MENF	87.55	81.43	84.38	73.55	74.14	73.85	57.87	71.45	63.95	74.06
SNF	88.14	83.21	85.60	74.44	76.29	75.35	61.57	73.15	66.86	75.93

In Table II we show the results on gold mentions, and compare with baseline system [24], Lassalle et al. [14], Xi et al. [6] and Lassalle & Denis [25]. Compared to baseline system and Xi et al. [6], our proposed system has improved CoNLL scores which are 4.76 and 5.61 improvement, respectively. The state-of-the-art approach [25] outperforms our proposed system, but they did not solve predication mention task. Instead, we evaluate our proposed model on both gold mention task and prediction mention task, which is shown in the following section.

TABLE II. PERFORMANCES COMPARISONS ON TEST SET WITH GOLD MENTION

System	MUC			B^3			CEAF			CoNLL
	P	R	F1	P	R	F1	P	R	F1	Mean
Lassalle et al. [14]	83.23	73.72	78.19	73.50	67.09	70.15	47.30	60.89	53.24	67.19
Xi et al. [6]	83.33	91.67	87.30	64.39	78.29	70.66	48.26	58.73	52.98	70.32
Lassalle & Denis [25]	88.40	87.04	87.71	78.49	78.14	78.31	77.93	81.92	79.88	81.97
Baseline [24]	87.11	71.43	78.49	83.69	60.77	70.41	69.58	60.32	64.62	71.17
This work (SNF)	88.14	83.21	85.60	74.44	76.29	75.35	61.57	73.15	66.86	75.93

Figure 4 shows that the CoNLL score distribution by sum of mention and antecedent span. The long spans of mention pairs perform better than the short span in our proposed model (using SNF), especially, spans above 20 have significant improvements compared with the baseline model.

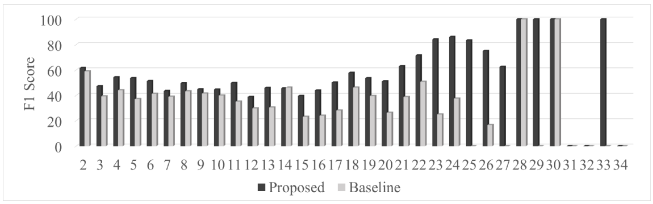

Figure 4. The F1 score distribution on proposed and baseline systems.

C. Results on Predicted Mention

In Table III we show the results with predicted mentions, and compare with baseline system [24], Durrett & Klein [26], Clark & Manning [5], Wiseman et al. [15], Wiseman et al. [16] and Clark & Manning [18]. On the unweighted mean score as CoNLL, our proposed system has 5.24% loss compared to Clark & Manning [18] proposed system. But our proposed model has the best 64.66 of B^3 score compared with all systems. The predicted mentions in our proposed system are detected by [24].

We can conclude the proposed system using predicted mentions on B3 score outperforms state-of-the-art approaches. In addition, it is safe to say that our proposed system is able to improve the coreference resolution for long spans significantly because CNN convolutes a set of words of long-span mention, which have rather complex meanings. Our proposed system has competitive performance compared

with other state-of-the-art systems and it can directly and intuitively learn string information to improve the performance of the mention ranking model. Our proposed system has not performed better than [18], but we have shown competitive performance on **B³**.

TABLE III. PERFORMANCES COMPARISONS ON TEST SET WITH PREDICTED MENTION

System	MUC			B^3			CEAF			CoNLL
	P	R	F1	P	R	F1	P	R	F1	Mean
Durrett & Klein [23]	72.61	69.91	71.24	61.18	56.43	58.71	56.17	54.23	55.18	61.71
Clark & Manning [5]	76.12	69.38	72.59	65.64	56.01	60.44	59.44	52.98	56.02	63.02
Wiseman et al. [15]	76.23	69.31	72.60	66.07	55.83	60.52	59.41	54.88	57.05	63.39
Wiseman et al. [16]	77.49	**69.75**	73.42	66.83	56.95	61.50	62.14	53.85	57.70	64.21
Clark & Manning [18]	**78.93**	**69.75**	**74.06**	70.08	56.98	62.86	**62.48**	**55.82**	**58.96**	**65.29**
Baseline [21]	65.54	63.50	64.51	**71.79**	54.46	61.94	46.28	53.34	49.56	58.67
This work (SNF)	69.64	65.05	67.26	71.45	**59.05**	**64.66**	50.96	45.76	48.22	60.05

V. CONCLUSION

We propose a deep learning framework for coreference resolution. Mention/antecedent representations are obtained by CNN to convolving word embeddings. The deep architecture for coreference resolution has effective performance compared with the baseline. The singleton classification increases the performance and the proposed system achieves a competitive performance compared with the state-of-the-arts. We also prove that CNNs indeed bring a certain level of benefits and able to improve the coreference resolution of long spans significantly. In the future, we aim to investigate other global information besides multiple dependency words of word pairs in the framework.

REFERENCES

[1] M. Recasens, M. C. de Marneffe, and C. Potts, "The life and death of discourse entities: identifying singleton mentions," Proc. the 2013 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013), Jun. 2013, pp. 627–633.

[2] Z. Chen, and H. Ji, "Graph-based event coreference resolution," Proc. the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4), Aug. 2009, pp. 54–57.

[3] Y. Versley, S. P. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang, A. Moschitti, "BART: A Modular Toolkit for Coreference Resolution", Proc. the 46 Annu. Meeting of the Assoc. for Computational Linguistics (ACL 2008), Jun. 2008, pp. 9-12.

[4] V. Ng, and C. Cardie, "Improving machine learning approaches to coreference resolution," Proc. the 40th Annu. Meeting of the Assoc. for Computational Linguistics (ACL 2002), July 2002 pp. 104–111.

[5] K. Clark, and C. D. Manning, "Entity-centric coreference resolution with model stacking," Proc. the 53th Annu. Meeting of the Assoc. for Computational Linguistics (ACL 2015), Jul. 2015, pp. 1405–1415.

[6] X. F. Xi, G. Zhou, F. Hu, and B. Fu, "A convolutional deep neural network for coreference resolution via modeling hierarchical features," Proc. the 2015 Int. Conf. on Intell. Sci. and Big Data Eng. (IScIDE 2015), Jun. 2015, pp. 361–372.

[7] T. Mikolov, W. T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," Proc. Proc. the 2013 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013), Jun. 2013, pp. 746–751.

[8] J. Pennington, R. Socher, and C. D. Manning, "Glove: global vectors for word representation," Proc. the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP 2014), Oct. 2014, pp. 1532–1543.

[9] Y. Kim, "Convolutional neural networks for sentence classification," Proc. the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP 2014), Oct. 2014, pp. 1746–1751.

[10] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," Proc. the 52th Annu. Meeting of the Assoc. for Computational Linguistics (ACL 2014), Jun. 2014, pp. 655–665.

[11] V. Stoyanov and J. Eisner, "Easy-first coreference resolution," Proc. the 23rd Int. Conf. on Computational Linguistics (COLING 2010), Aug. 2012, pp. 2519–2534.

[12] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A machine learning approach to coreference resolution of noun phrases," Computational Linguistic, vol. 27, jun. 2001, pp.521–544.

[13] E. Bengtson, and D. Roth, "Understanding the value of features for coreference resolution," Proc. the 2008 Conference on Empirical Methods on Natural Language Processing (EMNLP 2008), Oct. 2008, pp. 294–303.

[14] E. Lassalle, and P. Denis, "Improving pairwise coreference models through feature space hierarchy learning," Proc. the 51th Annu. Meeting of the Assoc. for Computational Linguistics (ACL 2013), Aug. 2013, pp. 497–506.

[15] S. Wiseman, A. M. Rush, S. M. Shieber, and J. Weston, "Learning anaphoricity and antecedent ranking features for coreference resolution," Proc. the 2015 Assoc. of Computational Linguistics (ACL 2015), Jul. 2015, pp. 92–100.

[16] S. Wiseman, A. M. Rush, and S. M. Shieber, "Learning global features for coreference resolution," Proc. the Human Language Technology and North American Assoc. for Computational Linguistics (NAACL-HLT 2016), Jun. 2016, pp. 994–1004.

[17] R. Stuckardt, "Applying backpropagation networks to anaphor resolution," Proc. the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007), Mar. 2007, pp. 107–124.

[18] K. Clark, and C. D. Manning, "Improving coreference resolution by learning entity-level distributed representations," Proc. the 2016 Assoc. of Computational Linguistics (ACL 2016), Aug. 2016, pp. 643–653.

[19] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "Conll-2012 shared task: modeling multilingual unrestricted coreference in ontonotes," Proc. the 2012 Joint Conf. on EMNLP and CoNLL-Shared Task, Jul. 2012, pp. 1–14.

[20] M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman, A model theoretic coreference scoring scheme. In Proc. the 6th conference on Message understanding, 1995, pp. 45–52.

[21] A. Bagga, and B. Baldwin, Algorithms for scoring coreference chains, In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Conference, 1998, pp. 563–566.

[22] X. Luo, On coreference resolution performance metrics, In Proc. the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 25–32.

[23] D. P. Kingma, and J. L. Ba, "Adma: A method for stochastic optimization," Proc. the ICLR 2015, May 2015, pp. 1–13.

[24] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. D. Manning, "A multi-pass sieve for coreference resolution," Proc. the 2010 Conference on Empirical Methods on Natural Language Processing (EMNLP 2010), 2010, pp. 492–501.

[25] E. Lassalle, and P. Denis, "Joint anaphoricity detection and coreference resolution with constrained latent structures," Proc. the 29th AAAI Conf. on Artificial Intell. (AAAI 2015), Jan. 2015, pp. 2274–2280.

[26] G. Durrett, and D. Klein, "A joint model for entity analysis: Coreference, typing, and linking," Transactions of the Assoc. for Computational Linguistics, vol. 2, 2014, pp. 477–490.