Deep Cross-Lingual Coreference Resolution for Less-Resourced Languages: The Case of Basque

Gorka Urbizu, Ander Soraluze and Olatz Arregi
Ixa group, University of the Basque Country (UPV/EHU)
gurbizu002@ikasle.ehu.eus
{ander.soraluze, olatz.arregi}@ehu.eus

Abstract

In this paper, we present a cross-lingual neural coreference resolution system for a less-resourced language such as Basque. To begin with, we build the first neural coreference resolution system for Basque, training it with the relatively small EPEC-KORREF corpus (45,000 words). Next, a cross-lingual coreference resolution system is designed. With this approach, the system learns from a bigger English corpus, using cross-lingual embeddings, to perform the coreference resolution for Basque. The cross-lingual system obtains slightly better results (40.93 F1 CoNLL) than the monolingual system (39.12 F1 CoNLL), without using any Basque language corpus to train it.

1 Introduction

Coreference resolution, the task of identifying and clustering all the expressions referring to the same real-world entity in a text, is essential in any Natural Language Processing (NLP) task that includes language understanding. For instance, tasks such as text summarisation (Steinberger et al., 2007), question answering (Vicedo and Ferrández, 2006), sentiment analysis (Nicolov et al., 2008) or machine translation (Werlen and Popescu-Belis, 2017) can benefit from coreference resolution.

In the last few years, we have witnessed how the revolution of neural networks and deep learning has improved the previous results in almost any NLP task. Big improvements in results were also obtained in coreference resolution in the last two years using neural approaches, mainly for English.

Although there is work in progress in languages other than English using neural networks, the results obtained are not so good in all of them. This is mostly due to smaller corpus sizes, which affects neural approaches negatively. The situation of less-resourced languages is even harder, as they have smaller datasets and annotating them is an arduous task to carry out by hand.

In this paper, we present a monolingual neural coreference resolution system for Basque. Subsequently, we try a cross-lingual approach to analyze whether it is possible to build a language independent coreference resolution system that obtains competent results when applied to less-resourced languages. To this end, we build a system which learns exclusively from an English corpus and apply it to resolve coreference in Basque texts. Afterwards, we compare the results of the monolingual system with a small dataset of the target language, and those of the cross-lingual system that learns from a bigger available corpus of another language, but not the target language.

The paper is organized as follows. Section 2 introduces related work. Section 3 describes the model built for coreference resolution. In section 4, we present the monolingual and cross-lingual experimental setups. Section 5 contains the obtained results. Finally, Section 6 presents our conclusions and future work.

2 Related Work

Coreference resolution has been handled with different techniques during the last few decades until deep learning techniques spread in the field. Among the most influential works are the rule-based system by H. Lee et al. (2013) and machine learning based systems by Soon et al. (2001) and Versley et al. (2008).

One of the first successful neural coreference resolution system (Wiseman et al.) obtained state-of-the-art results. Similar works followed, and although they differ in the method used for generating instances, all of them worked with automatic mentions and rule-based extracted features as input to a feedforward deep neural architecture (Clark and Manning; Wiseman et al., 2016).

The coreference resolution system that obtains

Proceedings of the 2nd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2019), pages 35–41,
Minneapolis, USA, June 7, 2019.

the best results in the current state-of-the-art is an end-to-end neural system, which is presented in K. Lee et al. (2017) and K. Lee et al. (2018). This system does not use any automatically preprocessed mentions or features, and it is able to find the needed features in the raw text, so it does not need any annotation other than the coreferential relations in the corpus. This manner, error propagation from the features extraction is reduced by learning those within the same neural network.

Neural coreference resolution systems for other languages have been created as well. For instance, in Clark and Manning they develop a system for Chinese, in Park et al. (2016) for Korean, and in Nitoń et al. (2018) for Polish.

Moreover, there has been some recent research to build cross-lingual systems for coreference resolution, as cross-lingual transfer learning has given good results in some other NLP tasks such as machine translation or language modeling (Lample and Conneau, 2019). Cruz et al. (2018) used neural networks to solve coreference for Portuguese by learning from Spanish, a related language, using cross-lingual word embeddings. Kundu et al. presented a similar system for Spanish and Chinese using English for training.

As regards the Basque language, this is the first work about neural coreference resolution. Nevertheless, a rule-based coreference resolution system (Soraluze et al., 2015) and a machine learning based system (Soraluze et al., 2016) have been developed. Both of which used a rule-based mention detector (Soraluze et al., 2017).

3 Model

In this section, the neural coreference resolution model, which is used for the experiments carried out, is presented.

The model used for coreference resolution for Basque is based on the neural system developed for Polish (Nitoń et al., 2018). After considering and discarting different models, it was chosen because both languages share some features such as being agglutinative or having free word order, and it obtained competitive results.

We use the mention-pair model to create instances, as in (Nitoń et al., 2018). They demonstrated that the mention-pair model obtains better results than the entity-mention for Polish.

In our case, gold mentions are used for training and development sets, and gold and automatic mentions are used for the test set, so we can see the effect of the performance of the mention detector in the results.

Once mention pairs are created, we extract some features of each mention and the mention pair to feed the neural network. In this work, we use pretrained 300-dimensional FastText embeddings (Bojanowski et al., 2017). They work with substring information, and this gives better results with morphologically rich languages such as Basque.

For each mention, we extract the following features:

- An average of the embeddings of the words that make up the mention (300 dimensions).
- An average of the embeddings of the words in the sentence in which the mention appears (300).

We extract the following features for the mention pair:

- Distance in words between the mentions, represented as binary features[1] (11).
- Distance in mentions between the mentions, represented as binary features (11).
- Whether mentions are in the same sentence (1).
- String matching (1).
- Lemma matching (1).
- Language[2]: Basque or English (1).

These features are easy to obtain for any language, and need very little preprocessing, just the lemmatization. In total, we obtain instances of 1,226 dimensions.

3.1 Neural Network

In this work, we use a fully connected network of 3 hidden layers, with 500, 300 and 100 neurons in each, and a single neuron in the output layer. The neural network takes instances of 1,226 dimensions in the input layer, and it returns a number between 0 and 1 in the output. The activation functions used are ReLU in the hidden layers and

[1] Binned into one of the following slots [0,1,2,3,4,5-7,8-15,16-31,32-63,64+,discontinuous].

[2] Included with the purpose of training on mixed language corpus in the future.

sigmoid in the output layer. ReLU function computes a positive number, while sigmoid function computes a number between 0 and 1.

$$\text{Input vector: } x = [e_i, e_j, e_{ij}]$$

$$\text{1st hidden layer: } h_1 = \text{ReLU}(W_1^T x + b_1)$$

$$\text{2nd hidden layer: } h_2 = \text{ReLU}(W_2^T h_1 + b_2)$$

$$\text{3rd hidden layer: } h_3 = \text{ReLU}(W_3^T h_2 + b_3)$$

$$\text{Output layer: } p(i,j) = sigmoid(w^T h_3)$$

Where e_i and e_j are the features of each mention, e_{ij} the features of the mention pair, W the weights and b the biases.

The neural network was trained to minimize the binary cross-entropy function. We trained the model for 2 epochs using a mini-batch size of 64. We used Adam optimization (Kingma and Ba, 2014), batch-normalization (Ioffe and Szegedy, 2015), and a dropout rate (Srivastava et al., 2014) of 0.2. The neural network was implemented using python library *KERAS*[3].

The mention pairs with a higher value than a threshold in the predictions are grouped in the same coreference cluster in testing time. To obtain the optimal threshold values, we used the development set.

4 Experimental Setup

Two experiments were carried out, both in similar conditions to be able to compare the outputs. In the first experiment, we trained the model described in the previous section with the available corpus of the Basque language for coreference. After that, we trained the system using a big corpus of English to see if the coreference resolution task could be learnt using transfer learning from another language.

4.1 Corpora

For the next experiments two corpora for coreference resolution are used, the EPEC-KORREF corpus (Ceberio et al., 2018) for Basque, the target language, and the OntoNotes English corpus (Hovy et al., 2006).

EPEC-KORREF[4] corpus is a Basque corpus, composed of news, of around 45K words and 12K mentions, which has mentions and coreferential relations, including singletons, annotated. The corpus is already divided into training, development and test sets, more details about the partition are shown in Table 1.

	Words	**Mentions**	**Clusters**	**Singletons**
Train	23,520	6,525	1,011	3,401
Dev	6,914	1,907	302	982
Test	15,949	4,360	621	2,445
Total	46,383	12,792	1,934	6,828

Table 1: EPEC-KORREF corpus

OntoNotes corpus is an English corpus with text from a variety of domains of more than one million words, with annotated mentions and coreferential relations. We used only newswire (nw), and broadcast news (bn) sets, avoiding conversation sets, in order to have texts of the same domain (around 825K words and 100K mentions). The details about the corpus are shown in Table 2.

	Words	**Mentions**
nw	625,000	75,000
bn	200,000	24,000

Table 2: OntoNotes corpus

4.2 Monolingual System

To develop the monolingual system, the neural model presented in Section 3 was trained on the EPEC-KORREF Basque corpus. In Figure 1, we can see how a train instance is generated from a coreferential mention pair of the following sentence:

Gaur egungo 15 herrialdeetatik 27ra igaro beharko du erdiko epera [Europar Batasunak], Europa ekialdeko eta hego ekialdeko 12 herrialde [bere] baitan hartuta.

"From the 15 countries of today, the [European Union] will have to change to 27 in the medium term, taking on [its] own 12 countries from west and southwest Europe."

The threshold for clustering mentions referring to the same entity was settled at 0.5 in the development set.

4.3 Cross-Lingual System

The same neural model presented in Section 3 is used to develop the cross-lingual system. However, in this case, it is trained on the English corpus, without using any corpus of the target lan-

[3]https://keras.io/

[4]http://ixa.si.ehu.es/node/4487

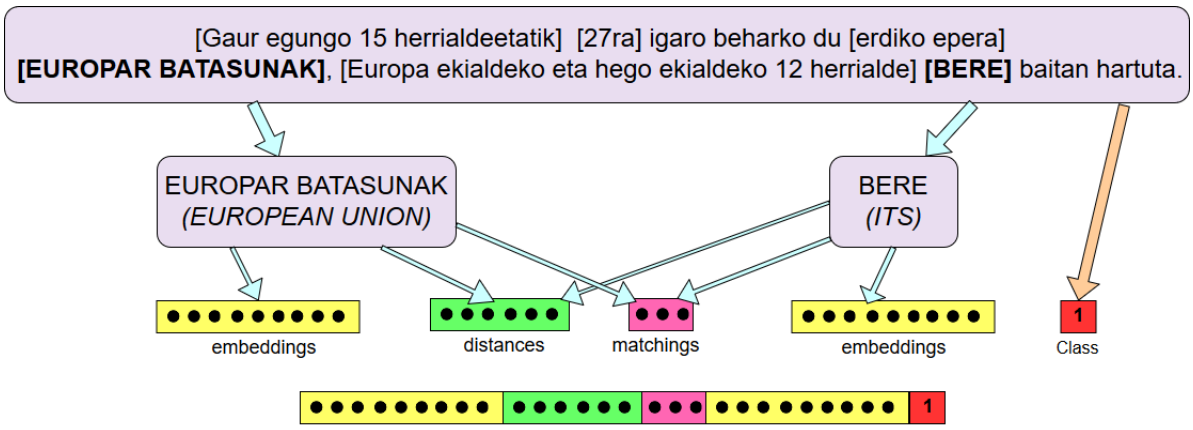

Figure 1: Example of an instance for a positive mention pair

guage, for the task of coreference resolution for Basque.

For this purpose, we use cross-lingual embeddings, as the language in the training set and the test set is different. We did this using the VecMap tool (Artetxe et al., 2018), which maps embeddings of one language to the other without using any bilingual dictionary.

The threshold for clustering coreferential mentions was settled at 0.9 in the development set.

5 Results

The coreference clusters obtained in the output of each experiment were evaluated with the official scorer proposed by Pradhan et al. (2014) for coreference resolution, and we also added the more recent LEA metric.

The main metrics used in the task are MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_m$ and $CEAF_e$ (Luo, 2005), BLANC (Recasens and Hovy, 2011), LEA (Moosavi and Strube) and CoNLL, which is the average of MUC, B^3 and $CEAF_e$ (Denis and Baldridge, 2009).

The results for the monolingual system and the cross-lingual system are shown in Table 4.

Our monolingual system obtains 39.12 F1 and 53.19 F1 for the CoNLL metric with automatic mentions and gold mentions respectively. The difference of using automatic (F1 = 73.79) or gold mentions is considerable (more than 14 points), which shows the importance of mention detection in the results. Furthermore, the low values for MUC metric stand out, which shows that the model does create a small number of coreference links.

Similar results were obtained for the cross-lingual system. The results for some metrics, such as MUC, decrease slightly, while the results for other metrics, such as LEA, increase a bit. Our cross-lingual system obtains 40.93 F1 for the CoNLL metric when automatic mentions are used and 54.46 F1 with gold mentions. We obtain better results with the cross-lingual system without using the target language corpus for the training than when using the small corpus available for Basque.

Moreover, to contextualize the results we obtained, in Table 3, we can see the results of the neural cross-lingual system in comparison with previous coreference resolution systems for Basque. The results obtained are lower than those obtained by previous rule-based (Soraluze et al., 2015) and ML-based (Soraluze et al., 2016) systems with the same corpus.

	CoNLL	
System	(auto)	(gold)
Rule-based	55.98	76.51
ML-based	54.21	73.94
Neural cross-lingual	40.93	54.46

Table 3: Comparison with previous systems for Basque

In Table 5 we can see an example of the type of mistakes in the output of our cross-lingual system. Key refers to gold annotation and response to the output of the system. Parentheses are used to mark mentions and numbers to tag coreference clusters. In the given example, we can see that the system has problems to link pronouns to the coreference cluster that they belong. This mistakes at solving pronominal coreference, are more common with neural and ML approaches than in rule-based sys-

System	MD	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	LEA	CoNLL
Monolingual (auto)	73.79	9.72	54.83	49.66	52.81	29.41	29.40	39.12
Cross-lingual (auto)		8.30	58.61	53.27	55.87	29.14	36.34	40.93
Monolingual (gold)	100	15.81	74.60	63.10	69.17	53.28	39.87	53.19
Cross-lingual (gold)		10.00	79.90	68.09	73.47	51.91	49.30	54.46

Table 4: Results of monolingual and cross-lingual systems for gold and automatic mentions

Key	... eta (bera)$_1$, ((zailtasun hori)$_2$ gainditu duen munduko lehen emakumea)$_1$.
Response	... eta (bera)$_1$, (zailtasun hori)$_2$ gainditu duen (munduko lehen emakumea)$_3$.
Translation	... and (she)$_1$ is (the first woman in the world to overcome (that difficulty)$_2$)$_1$.

Table 5: Example of mistakes in the output

tems. Training our cross-lingual system on English might make this even harder, as Basque has gender-neutral pronouns and it is quite common to drop pronouns at subject or object positions.

6 Conclusions and Future Work

We present a neural coreference resolution system for Basque, and a cross-lingual system, which is trained on a bigger English corpus.

The results obtained with both systems are significantly lower than those obtained by previous non-neural systems for Basque. The results of the cross-lingual system (40.93 F1 CoNLL) are slightly better than the monolingual ones (39.12 F1 CoNLL), and this was obtained without using any target language corpus in the training phase.

Furthermore, we conclude that the corpus for Basque, of 45,000 words, is too small for a monolingual neural approach. Thus, the results obtained with the cross-lingual system are outstanding, as they improved the results obtained without using any corpus of the target language.

An in-depth error analysis needs to be done to understand better the results of both systems. Moreover, training the same model for coreference resolution for English would help to see whether the results obtained were due to the neural architecture and the model, or the small corpus and the cross-lingual approach. In addition, it might be interesting to see what results we would obtain with a simpler model, mostly for the monolingual system.

The cross-lingual approach needs to be investigated further. We are planning to apply this cross-lingual approach to the state-of-the-art neural network architecture (K. Lee et al., 2018), which might learn better, and could help to close the gap between results obtained with automatic and gold mentions.

Finally, this cross-lingual system could be tested for different language pairs, to see what language pairs give better results, with the aim of building a universal coreference resolution system, which would learn the task for many languages and resolve coreference for any other language.

Acknowledgments

This research was partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE; PROSA-MED project, TIN2016-77820-C3-1-R) and by the European Commission (LINGUATEC project, EFA227/16).

We thank the three anonymous reviewers whose comments and suggestions contributed to improve this work.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 79–85. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Klara Ceberio, Itziar Aduriz, Arantza Díaz de Ilarraza, and Ines Garcia-Azkoaga. 2018. Coreferential relations in Basque: the annotation process. *Journal of psycholinguistic research*, 47(2):325–342.

Kevin Clark and Christopher D Manning. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. 2018. Exploring Spanish corpora for Portuguese coreference resolution. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 290–295. IEEE.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42:87–96.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*.

Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 448–456.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Gourab Kundu, Avi Sil, Radu Florian, and Wael Hamza. Neural cross-lingual coreference resolution and its application to entity linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

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, 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.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *NAACL-HLT (2)*, pages 687–692. 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.

Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Nicolas Nicolov, Franco Salvetti, and Steliana Ivanova. 2008. Sentiment Analysis: Does Coreference Matter? In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, pages 37–40.

Bartłomiej Nitoń, Paweł Morawiecki, and Maciej Ogrodniczuk. 2018. Deep neural networks for coreference resolution for Polish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, pages 395–400.

Cheoneum Park, KyoungHo Choi, Changki Lee, and Soojong Lim. 2016. Korean Coreference Resolution with Guided Mention Pair Model Using Deep Learning. *ETRI Journal*, 38(6):1207–1217.

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, Baltimore, Maryland. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Ander Soraluze, Olatz Arregi, Xabier Arregi, and Arantza Diaz de Ilarraza. 2015. Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque. In *Procesamiento del Lenguaje Natural*, volume 55, pages 23–30.

Ander Soraluze, Olatz Arregi, Xabier Arregi, and Arantza Díaz de Ilarraza. 2017. Improving mention detection for Basque based on a deep error analysis. *Natural Language Engineering*, 23(3):351–384.

Ander Soraluze, Olatz Arregi, Xabier Arregi, Arantza Diaz de Ilarraza, Mijail Kabadjov, and Massimo Poesio. 2016. Coreference Resolution for the Basque Language with BART. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 67–73.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Jeek. 2007. Two Uses of Anaphora Resolution in Summarization. *Information Processing and Management*, 43(6):1663–1680.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pages 9–12. Association for Computational Linguistics.

Jose Vicedo and Antonio Ferrández. 2006. Coreference In Q&A. In *Advances in Open Domain Question Answering*, volume 32 of *Text, Speech and Language Technology*, pages 71–96. Springer.

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.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Using coreference links to improve Spanish-to-English machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40.

Sam Wiseman, Alexander M Rush, Stuart Shieber, and Jason Weston. Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. 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)*.

Sam Wiseman, Alexander M Rush, and Stuart M Shieber. 2016. Learning Global Features for Coreference Resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004.