Neural Coreference Resolution with Limited Lexical Context and Explicit Mention Detection for Oral French

Loïc Grobol
Lattice CNRS, 1 rue Maurice Arnoux, 92120 Montrouge, France
ALMAnaCH, Inria, 2 rue Simone Iff, 75589 Paris, France
loic.grobol@inria.fr

Abstract

We propose an end-to-end coreference resolution system obtained by adapting neural models that have recently improved the state-of-the-art on the OntoNotes benchmark to make them applicable to other paradigms for this task. We report the performances of our system on ANCOR, a corpus of transcribed oral French — for which it constitutes a new baseline with proper evaluation.

1 Introduction

In the last few years, coreference resolution systems based on artificial neural networks architectures have received much attention by tremendously improving upon the previous state-of-the-art. In particular, the system introduced by K. Lee et al. (2017) and refined in (K. Lee et al. 2018) have proved that relatively high scores could be achieved without relying on rich features and preprocessing pipelines.

However, these results were obtained in the paradigm of the CoNLL-2012 shared task (Pradhan et al. 2012) and it is not self-evident that they are generalisable to other datasets, other domains and other languages. For instance, the choice in to not include singleton mentions in the CoNLL-2012 dataset is quite uncommon and might rightfully be suspected to affect the evaluation of coreference resolution architectures (see for instance the comparisons made by Poesio et al. (2018)).

In this work, we present an adaptation of K. Lee et al. (2018)'s system (henceforth E2EC[1]) to make it more suitable to other paradigms. We evaluate our system on ANCOR (Muzerelle et al. 2014) — a corpus of transcribed oral French.

[1]From its official repository https://github.com/kentonl/e2e-coref.

2 Related Works

There is a large existing body of work on coreference resolution spanning from the 1970s of which Poesio et al. (2016) provides an exhaustive review. In recent years, the field has been dominated by machine learning approaches — with the notable exception of the rule-based system of H. Lee et al. (2013) — from shallow learning approaches (C. Ma et al. 2014; Björkelund and Kuhn 2014; Durrett and Klein 2014) to systems based on artificial neural network architectures (Clark and Manning 2016a; Clark and Manning 2016b; Wiseman et al. 2015; Wiseman et al. 2016), gradually reducing their dependency on rich features coming from preprocessing pipelines using linguistic knowledge. One of the last incarnations of this tendency is the E2EC system introduced by K. Lee et al. (2017), which has close to no dependency to external resources (except for pretrained word embeddings derived from non-annotated data) and yet reaches state-of-the-art performance on the most common benchmark: the fully end-to-end track of the CoNLL-2012 shared task (Pradhan et al. 2012).

At the core of E2EC is the idea of performing coreference detection on the set of all possible text spans instead of using markables detected by an independent mention detector. This is made possible through the use of dense representations of arbitrary text spans derived from the internal states of recurrent neural networks. K. Lee et al. (2018) introduced further improvements to this model, most notably a higher-order approach to coreference detection using incremental refinements of its spans representations based on their antecedent distributions and an early pruning of antecedent candidates based on a coarse-to-fine scoring strategy.

However, to the best of our knowledge, using a simple classifier on these span embeddings to detect mentions had not yet been explored. Even

Proceedings of the 2nd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2019), pages 8–14,
Minneapolis, USA, June 7, 2019.

Sanh et al. (2018) — which used the AllenNLP (Gardner et al. 2018) implementation of E2EC for the coreference detection part of its system — used a sequence labelling-based model for entity-mention detection instead.

On our target corpus, ANCOR (Muzerelle et al. 2014), there have been relatively few works focused on automatic coreference resolution. Désoyer et al. (2015) presented an exploration of shallow learning techniques for the coreference detection phase, using the rich features provided by the gold annotations, delegating to further works the task of automatically detecting these features for a full-end-to-end pipeline. Some exploratory work on detecting mentions and these features has been presented in Grobol et al. (2017) with encouraging but limited results. The independent work presented by Godbert and Favre (2017) treated coreference resolution with a rule-based system on top of the MACAON pipeline (Nasr et al. 2011), focusing on pronominal anaphora resolution, yet reaching encouraging overall performances.

3 Model

Our architecture is mostly an adaptation of the version of E2EC presented by K. Lee et al. (2018), modified to address the difficulty of applying it to other paradigms, which is mainly due to two factors. The first one is that E2EC always operate at the level of a whole document. In principle, this would be a desirable property, since coreference chains are document-level objects. However, during the training process, it implies that the whole document has to be kept in memory and that error backpropagation must span all of its processing, which results in impractical memory and computing requirements. K. Lee et al. (2017) address this by performing a variety of aggressive pruning at every step, which complexifies its implementation and makes the training process less efficient. Despite this, the final implementation is still quite demanding in resources, particularly with huge documents and not necessarily effective on data — like ANCOR — where the context outside of the immediate vicinity of a span might be very noisy. It also prevents the use of common training techniques, like mini-batching and sample shuffling, since it imposes the use of batches that are each the size of a whole document.

The second characteristic we address is the lack of explicit mention detection. E2EC does not make a distinction between non-mention text spans and singleton mentions and as such, does not actually perform mention detection[2]. This is not a real problem on CoNLL-2012, but it is one for corpora that include singleton mentions. It also prevents the use of gold mentions to evaluate the actual coreference detection capabilities of a system without the bias induced by mention detection.

To alleviate these issue, our system are then should only take into account the immediate context of text spans rather than whole documents and that perform mention detection as an explicit step in order to take singleton mentions into account. In addition to these adaptations, we also added a certain number of incremental modifications inspired from recent works on sequence embeddings in neural networks. These modifications were added during our initial experiments on the mention detection part, for which they improved the global scores on the development dataset, but at the time of writing, we did not assess their actual impact on the whole architecture.

Words representations Similarly to e.g. X. Ma and Hovy (2016), we use a combination of pretrained word embeddings and character-level encodings derived from a recurrent neural layer (in our case a bidirectional GRU (Cho et al. 2014)), which helps with noisy inputs (including disfluencies, incomplete words and typos in ANCOR) but also unknown words and casing information that is not available to the pretrained word embeddings.

Span embeddings The span embeddings are computed using a combination of recurrent and self-attentional mechanisms. At the core is a bidirectional LSTM with two layers, that we run on the sequence of the representations $(w_{-\ell}, \dots, w_0, \dots, w_{n-1}, w_n, \dots, w_{n+p})$ of the words of the span (from w_0 to w_{n-1}) and its immediate left and right contexts. We keep the hidden states $h_i = [\overleftarrow{h_i}, \overrightarrow{h_i}]$ of both directions of the top LSTM layer, and use them in three subsequent treatments

- The hidden states of the first and the last word of the span are kept as a pure recurrent representation $r = [h_0, h_{n-1}]$
- The self-attention soft-head mechanism introduced by K. Lee et al. (2017) is applied to the sequence $([w_0, h_0], \dots [w_{n-1}, h_{n-1}])$ with

[2] It does compute a "mention score", but more as way to reduce the computational complexity of the architecture than as an explicit mention detection, and the correlation between this score and "mentionity" of text spans has not yet been studied.

two separate heads (inspired by the multi-head attention mechanism of Vaswani et al. (2017)) whose concatenation gives us an attentional representation a

- The final states of the LSTM are kept as a representation of the span context $c = [\overrightarrow{h_{-\ell}}, \overleftarrow{h_{n+p}}]$. This was not part of the initial model, but we found that it helps significantly (at least for mention detection) on the most interactive parts of ANCOR.

The final span embedding s is then obtained by concatenating these three representations and f, a low-dimension feature embedding that encodes the length of the span and passing the result through a feedforward network giving $s = \mathrm{FFNN}_{\mathrm{out}}(r, a, c, f)$.

Mentions detection The mention detecting layer is a simple feedforward classifier that takes s as input and outputs a vector of class scores: "None" for non-mention spans and depending on the corpus, either a simple "NP" class for all mentions or distinct classes for noun phrases and pronouns.

Antecedents scoring The antecedent scoring layer assign coreference scores to mention/antecedent pairs using the same coarse-to-fine second-order inference mechanism as E2EC, with the representation refining done solely for the mention and not its antecedents. The only other variation is that instead of fixing the score of the dummy antecedent ε for a span s to 0 we instead compute a specific mention-new score by applying a simple feedforward network on s. This was motivated by the higher number of non-anaphoric mentions in ANCOR (again due to the inclusion of singleton mentions) and seems to affect the final coreference scores positively, although a more formal assessment of this is still needed.

4 Evaluation

Following the recommendations of Recasens (2010, p. 122) and Salmon-Alt et al. (2004) we evaluate our system separately on the two subtasks that it performs. For mention detection, we report the usual Precision, Recall and F-score detection metrics. For coreference resolution, we use the CoNLL-2012 metrics (Pradhan et al. 2014) including BLANC (Recasens and Hovy 2011). This is a standard evaluation procedure for coreference resolution systems — as seen for example in the CRAC18 shared task (Poesio et al. 2018). It also allows us to compare our system with other works on ANCOR (Désoyer et al. 2015; Godbert and Favre 2017) and to assess the actual capabilities of our antecedent scoring module by avoiding the noise caused by the inevitable mention detection errors.

5 Experiments

5.1 Data

The primary object of our study is the ANCOR corpus (Muzerelle et al. 2014). ANCOR is, for now, the only currently publicly available[3] corpus of French with coreference annotations whose size is sufficient for machine learning purposes, with around 418 000 words. The source materials of this corpus are *speech transcriptions*[4] , in most part long interviews with low interactivity taken from the ESLO corpus (Baude and Dugua 2011) and smaller parts with higher interactivity[5]. Its annotations include coreference and morphosyntactic annotations for noun phrases and pronouns including singleton mentions, but no linguistic annotations of other elements.

Since existing works on ANCOR do not provide detailed training/development/test partitions, ours is probably different, but we tried to stay reasonably close to the one described by Désoyer et al. (2015), with about 60 % of the corpus devoted to the training set. However, we chose to keep most of the rest to the test set, in order to provide more significant final scores. The final distribution is 59 %/12 %/29 %, with a fairly homogeneous distribution of the different subcorpora, in order to minimize the disparities caused by their various levels of interactivity and topics.

5.2 Hyperparameters

In order to stay close to the original E2EC model, we have mostly kept the same hyperparameters and mention here only those that we changed. All of these changes were motivated by purely empirical observations of the performance of the model on the ANCOR development set.

[3] Another large scale corpus exists (Tutin et al. 2000) but is not publicly available.

[4] The fact that the source material is not written (or controlled oral) language — as in most coreference corpora — is another factor that might skew the comparison with other works, but assessing its actual impact would require a comparable corpus for written French, which does not exist yet.

[5] See Brassier et al. (2018) for details on this part.

Table 1: Coreference resolution

System	MUC			B^3			$CEAF_e$			CoNLL	BLANC		
	P	R	F	P	R	F	P	R	F	Avg.	P	R	F
Désoyer et al. (2015)	—	—	63.5	—	—	83.8	—	—	79.0	75.3	—	—	67.4
Godbert and Favre (2017)	—	—	—	—	—	—	—	—	—	—	—	—	65.7[1]
Our model[2]	72.3	47.7	57.3	89.7	71.0	79.2	72.8	86.0	79.4	72.0	78.2	60.1	65.7

[1] It is not clear if the score reported as BLANC by Godbert and Favre (2017) actually takes into account both coreference and non-coreference links after rebuilding mention clusters or is simply the raw F-score of the antecedent finder.

[2] Averages on 5 runs.

Words representations We use word embeddings pretrained on the Common Crawl for FastText (Grave et al. 2018) and fine-tuned during training. The character embeddings are not pretrained and are initialized randomly.

Span encoding The span contexts considered are of size 10 on both sides. We only consider spans of at most 25 words to reduce the time and material requirements. Experiments made with longer spans did not show significantly different results. Our hypothesis is that too few mentions are longer than this limit to impact the learning.

Antecedent scoring During the antecedent scoring phase, only the 100 previous mentions are considered for coarse scoring and only the 25 best-scoring antecedents are kept for fine-scoring.

Training We trained the network sequentially, first on mention detection, then on antecedent scoring. For both, the trainable parameters were optimized using the AdamW (Loshchilov and Hutter 2019) optimizer.

For mention detection, we minimize the class-weighted cross-entropy (Panchapagesan et al. 2016) with a weight of 1 for "None" and 3 for the mention span class. We also undersample the spans in the training set to a maximum ratio of 90 % of non-mention spans, to alleviate the usual issues of neural classifiers with severe classes imbalance. For antecedent scoring, we follow K. Lee et al. (2017) and optimize the sum of the log-likelihood of all the correct antecedents of each mention.

5.3 Results

Mention detection Table 2 presents the results of our experiments with mention detection compared to the baseline of Grobol et al. (2017) — which consists in merely extracting all the NP from the output of an off-the-shelf parser — and the performance reported by Godbert and Favre (2017). Considering the sparsity of its own resources, our system does not fare too bad, even though its precision shows a lot of room for improvements.

Table 2: Mention detection

System	P	R	F
Grobol et al. (2017)	57.28	77.07	65.72
Godbert and Favre (2017)	90.05	87.86	88.94
Our model[1]	82.99	89.07	85.87

[1] Averages on 5 runs.

Coreference resolution Table 1 presents the performances of our system for coreference resolution and compare it with those of previous works. Note that we didn't compare with the performances of the original E2EC on ANCOR, since there is no simple way to provide it with gold mentions[6] at either training or test time, nor to make it distinguish between singleton mention and non-mention spans without significantly modifying it.

As mentioned in the previous sections, the existing work on ANCOR have been developed in different paradigms and as such are not entirely comparable to ours. This is particularly true for Désoyer et al. (2015), which relies on gold features, and as such was able to get very high scores on all metrics with a relatively simple system, these results should thus be considered as an upper baseline than a real benchmark. In addition, none of these works report the full detailed CoNLL-2012 metrics, which limits the interpretability of these results. Taking these reserves into account the performances of

[6] In the usual sense and not in the "anaphoric gold mentions" sense used in K. Lee et al. (2017).

our system suggests that neural architectures can indeed be effective in the paradigm of ANCOR.

6 Conclusion

We presented an end-to-end coreference resolution system inspired by the most recent models to reach state-of-the-art performance on the classic CoNLL-2012/Ontonotes dataset. Our system is made suitable for experiments on other datasets by the extraction of an explicit mention detection phase from the original end-to-end architecture of K. Lee et al. (2017) and the restriction of the input representations to the immediate contexts of the markables. Given these adaptations, we report performances on ANCOR — a corpus of transcribed oral French— that are close to those reported by previous works, which required the use of considerably more linguistic knowledge.

This tends to prove that knowledge-poor, end-to-end neural architectures are applicable to coreference detection tasks beyond the OntoNotes benchmark. It also provides future works on coreference resolution for French with a baseline for full evaluations on both parts of the task.

However, our system has only been tested on a single corpus so far, and its architecture is optimized for it. Further assessment of its capabilities should include further tests on other, comparable, corpora such as ARRAU (Poesio and Artstein 2008), the Polish Coreference Corpus (Ogrodniczuk et al. 2016) or the upcoming DEMOCRAT corpus (Landragin 2016). Proper evaluation should also eventually include comparisons on the CoNLL-2012 dataset itself, possibly in the "gold mention boundaries" settings for a better comparability.

Acknowledgements

This work is part of the "Investissements d'Avenir" overseen by the French National Research Agency ANR-10-LABX-0083 (Labex EFL), and is also supported by the ANR DEMOCRAT (Describing and Modelling Reference Chains: Tools for Corpus Annotation and Automatic Processing) project ANR-15-CE38-0008.

References

Olivier Baude and Céline Dugua. 2011. (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? *Corpus*. Varia, 10, 2011: 99–118.

Anders Björkelund and Jonas Kuhn. 2014. Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-Local Features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 47–57. Volume 1. Association for Computational Linguistics.

Maëlle Brassier, Alexis Puret, Augustin Voisin-Marras, and Loïc Grobol. 2018. Classification par paires de mention pour la résolution des coréférences en français parlé interactif. In *Actes de la Conférence jointe CORIA-TALN-RJC 2018*. Association pour le Traitement Automatique des Langues. Rennes, France.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016a. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2256–2262.

Kevin Clark and Christopher D. Manning. 2016b. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Volume 1. Association for Computational Linguistics. Berlin, Deutschland.

Adèle Désoyer, Frédéric Landragin, Isabelle Tellier, Anaïs Lefeuvre, and Jean-Yves Antoine. 2015. Coreference Resolution for Oral Corpus: a machine learning experiment with ANCOR corpus. *Traitement Automatique des Langues*. Traitement automatique du langage parlé, 55.2, May 2015: 97–121.

Greg Durrett and Dan Klein. 2014. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *Transactions of the Association for Computational Linguistics*, 2.0, Nov. 1, 2014: 477–490.

Matt Gardner et al. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform.

In *Proceedings of Workshop for NLP Open Source Software*, pages 1–6. Association for Computational Linguistics.

Elisabeth Godbert and Benoît Favre. 2017. Détection de coréférences de bout en bout en français. In *Actes de la 24e Conférence sur le Traitement Automatique des Langues Naturelles*. Association pour le Traitement Automatique des Langues. Orléans, France.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*. European Language Resource Association.

Loïc Grobol, Isabelle Tellier, Éric De La Clergerie, Marco Dinarelli, and Frédéric Landragin. 2017. Apports des analyses syntaxiques pour la détection automatique de mentions dans un corpus de français oral. In *Actes de la 24e Conférence sur le Traitement Automatique des Langues Naturelles*. Association pour le Traitement Automatique des Langues. Orléans, France.

Frédéric Landragin. 2016. Description, Modélisation et Détection Automatique Des Chaînes de Référence (DEMOCRAT). *Bulletin de l'AFIA*, 92, 2016: 11–15.

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, Dec. 2013: 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. Association for Computational Linguistics. København, Danmark.

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*, pages 687–692. Volume 2. Association for Computational Linguistics. New Orleans, Louisiana.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. arXiv: 1711.05101.

Chao Ma, Janardhan Rao Doppa, J. Walker Orr, Prashanth Mannem, Xiaoli Fern, Tom Dietterich, and Prasad Tadepalli. 2014. Prune-and-Score: Learning for Greedy Coreference Resolution. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2115–2126. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-End Sequence Labeling via Bi-Directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1064–1074. Volume 1. Association for Computational Linguistics. Berlin, Deutschland.

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurore Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 2014. ANCOR Centre, a Large Free Spoken French Coreference Corpus: Description of the Resource and Reliability Measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. European Language Resources Association. Reykjavík, Ísland.

Alexis Nasr, Frédéric Béchet, Jean-François Rey, Benoit Favre, and Joseph Le Roux. 2011. MACAON : An NLP Tool Suite for Processing Word Lattices. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 86–91. United States.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2016. Polish Coreference Corpus. In Zygmunt Vetulani, Hans Uszkoreit, and Marek Kubis, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 215–226. Lecture Notes in Computer Science. Springer International Publishing.

Sankaran Panchapagesan, Ming Sun, Aparna Khare, Spyros Matsoukas, Arindam Mandal, Björn Hoffmeister, and Shiv Vitaladevuni. 2016. Multi-Task Learning and Weighted Cross-Entropy for DNN-Based Keyword Spotting. In *Proceedings of Interspeech*, pages 760–764.

Massimo Poesio and Ron Artstein. 2008. Anaphoric Annotation in the ARRAU Corpus. In *Proceedings of the International Confer-*

ence on Language Resources and Evaluation. Marrakech, Morocco.

Massimo Poesio, Ron Stuckardt, and Yannick Versley. 2016. *Anaphora Resolution: Algorithms, Resources, and Applications*. Theory and Applications of Natural Language Processing. Springer Berlin Heidelberg.

Massimo Poesio et al. 2018. Anaphora Resolution with the ARRAU Corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22. Association for Computational Linguistics. New Orleans, Louisiana.

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*, pages 30–35. Volume 2. Association for Computational Linguistics. Baltimore, Maryland.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the Joint Conference on EMNLP and CoNLL*, pages 1–40. Association for Computational Linguistics. Jeju Island, Korea.

Marta Recasens. 2010. Coreference: Theory, Annotation, Resolution and Evaluation. Universitat de Barcelona, 2010.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand Index for Coreference Evaluation. *Natural Language Engineering*, 17.4, Oct. 2011: 485–510.

Susanne Salmon-Alt, Laurent Romary, Andrei Popescu-Belis, and Loïs Rigouste. 2004. Online Evaluation of Coreference Resolution. In *4th International Conference on Language Resources and Evaluation*. European Language Resources Association. Lisboa, Portugal.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2018. A Hierarchical Multi-Task Approach for Learning Embeddings from Semantic Tasks, Nov. 14, 2018: arXiv: `1811.06031 [cs]`.

Agnès Tutin, François Trouilleux, Catherine Clouzot, Éric Gaussier, Annie Zaenen, Stéphanie Rayot, and Georges Antoniadis. 2000. Annotating a Large Corpus with Anaphoric Links. In *Proceedings of the Third International Conference on Discourse Anaphora and Anaphora Resolution*. United Kingdom.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc. Long Beach, California.

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. Association for Computational Linguistics.

Sam Wiseman, Alexander M. Rush, Stuart M. Shieber, and Jason Weston. 2015. 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 of the Asian Federation of Natural Language Processing*, pages 1416–1426. Volume 1. Beijing, China.