A Simple Unsupervised Approach for Coreference Resolution using Rule-based Weak Supervision

Anonymous ACL submission

Abstract

Labeled data for the task of Coreference Resolution is a scarce resource, requiring significant human effort. While state-of-the-art coreference models rely on such data, we propose an approach that leverages an end-to-end neural model in settings where labeled data is unavailable. Specifically, using weak supervision, we transfer the linguistic knowledge encoded by Stanford's rule-based coreference system to the end-to-end model, which jointly learns rich, contextualized span representations and coreference chains. Our experiments on the English OntoNotes corpus demonstrate that our approach effectively benefits from the noisy coreference supervision, producing an improvement over Stanford's rule-based system (+3.7 F_1) and outperforming the previous best unsupervised model (+0.9 F_1). Additionally, we validate the efficacy of our method on two other datasets: PreCo and Litbank (+2.5 and +4 F_1 on Stanford's system, respectively).

1 Introduction

Coreference resolution is an important problem in language understanding. In the recent years, significant progress has been made on this task with coreference annotated corpora (Hovy et al., 2006) and deep neural network architectures (Wiseman et al., 2015; Clark and Manning, 2016a,b; Lee et al., 2017). Further gains have been obtained by leveraging contextualized text encoders like ELMo (Lee et al., 2018), BERT, SpanBERT, and Longformer (Kantor and Globerson, 2019; Joshi et al., 2019, 2020; Wu et al., 2020; Kirstain et al., 2021).

The progress in supervised coreference resolution has not been accompanied by analogous improvements in unsupervised methods. The best performing work in this domain is the unsupervised mention-ranking systems proposed by Ma et al. (2016). Approaches that do not rely on gold annotation are highly desirable for this task, as coreference corpora are expensive to create. Addressing this issue, weak supervision has been used for multilingual coreference resolution to automatically obtain labels for languages with no annotated datasets (Wallin and Nugues, 2017).

In this paper, we introduce a simple yet effective approach for unsupervised coreference resolution, which leverages an end-to-end span-ranking coreference model (Lee et al., 2018) and contextualized span representations. The end-to-end model is trained with weak supervision from Stanford's coreference system (Lee et al., 2011), which, in turn uses a set of linguistic rules for coreference. Previous works have used Stanford system's rules as feature extractors (Fernandes et al., 2012; Wiseman et al., 2015; Ma et al., 2016). However, our approach uses Stanford's rule-based sieves to produce noisy labels that are subsequently used to train the neural end-to-end resolver.

The rationale behind the use of Stanford's resolver for producing noisy labels lies in its ease of use and its modular structure, which allows us to interpret the value of the linguistic knowledge encoded in the system. Linguists building a coreference resolver in a new domain can encode their prior knowledge via rules and improve the Stanford system. Our approach would further boost the resolver by incorporating pre-trained representations. Nevertheless, our framework can be applied in combination with any method able to produce informative coreference labels.

We assess our approach on three coreference corpora: English OntoNotes (Pradhan et al., 2012), PreCo (Chen et al., 2018), and Litbank (Bamman et al., 2020). Our experiments show that the imperfect information contained in the noisy labels can be effectively used to train the end-to-end model, producing an improvement over Stanford's system. Experimenting with different pre-trained language models, we observe that using BERT boosts the performance of the end-to-end resolver. Results

further improve by using SpanBERT (Joshi et al., 2020), which outperforms previous unsupervised models (Ma et al., 2016) on the English OntoNotes benchmark. We also evaluate the approach on two other coreference datasets: PreCo and Litbank, and show strong gains over the Stanford system. Finally, we present a set of analyses that examine the information incorporated by weakly supervised training.

2 Method

Our approach relies on the *c2f-coref* end-to-end architecture proposed by Lee et al. (2018), and on the classic rule-based Stanford coreference system (Lee et al., 2011, 2013) for the CoNLL 2011 shared task (Pradhan et al., 2011).

Overview of c2f-coref The end-to-end coreference resolution system (Lee et al., 2017) uses a span-based neural model that learns a distribution $P(\cdot)$ over antecedents y for each span i. Spans are represented using fixed-length embeddings obtained via bidirectional LSTMs (Hochreiter and Schmidhuber, 1997) and taken as input by a pairwise scoring function.

Subsequent models revisited this approach: Lee et al. (2018) proposed the c2f-coref method, introducing coarse-to-fine antecedent pruning and embedding representations from ELMo (Peters et al., 2018) at the input to the LSTMs. Later, Joshi et al. (2019) used BERT to represent spans, demonstrating the power of pre-trained language models for coreference resolution. Most recently, Joshi et al. (2020) introduced SpanBERT and further improved the state of the art.

Stanford's Rule-based System Stanford's system is a deterministic coreference resolver consisting of a set of sieves applied in a cascade fashion. Initially, the *Mention Detection* considers all noun phrases, pronouns, and named entity mentions as candidate mentions, then filters them according to a set of exclusion rules. Specifically, each identified mention is considered as a singleton cluster. Then, akin to agglomerative clustering, the clusters are sequentially processed by the sieves. Each sieve embodies a specific linguistic rule and builds on the result of the previous sieve by merging a mention into a partially-formed entity cluster, depending on whether it satisfies a set of constraints. The architecture guarantees that high-precision constraints are given high priority (e.g., exact string match, head match), while rules with lower precision but higher recall are applied later (e.g., the Pronominal Coreference Sieve). We provide a description of the most important sieves in Appendix A.

Weak Supervision using Linguistic Rules Although Stanford's sieve-based system is unsupervised, it captures rich, task-specific coreference information in English, and we hypothesize that it could effectively serve as supervision for training the neural span-ranking model. By exploiting contextualized span representations within the end-to-end learning framework, the neural model can exhibit stronger generalization capabilities.

Specifically, we employ Stanford's system to obtain cluster labels, representing a *noisy* (i.e., non-gold) signal for both mention identification and coreference. As in the supervised case, only clustering information is observed. The training is carried out by optimizing the marginal log-likelihood of the antecedents $\tilde{y}$ implied by the noisy cluster assignment:

$$\log \prod_{i=1}^{N} \sum_{\tilde{y} \in \mathcal{C}(i)} P(\tilde{y})$$

where N is the total number of mentions in the document and $\mathcal{C}(i)$ is the set of antecedents of span i that are coreferent to i according to the cluster assignment produced by Stanford's system.

3 Experiments

We assess the proposed approach on three datasets: the English OntoNotes v5.0 data from the CoNLL-2012 shared task (Pradhan et al., 2012), PreCo (Chen et al., 2018), and Litbank (Bamman et al., 2020). We evaluate the c2f-coref model combined with different pre-trained language models (ELMo, BERT, and SpanBERT). These results are compared to the ones produced by Stanford's system, in order to show the efficacy of the noisy supervision. Moreover, we examine the performance of our weakly-supervised approach in contrast to two previous unsupervised models: Multigraph (Martschat, 2013) and the EM-based ranking model by Ma et al. (2016).

3.1 Experimental Setup

We use the original implementations of the ELMo-based c2f-coref[1] (Lee et al., 2018) and of the BERT/SpanBERT-based models[2] (Joshi et al.,

[1] https://github.com/kentonl/e2e-coref

[2] https://github.com/mandarjoshi90/coref

	MUC			B^3			$CEAF_{\phi_4}$			CoNLL
	P	R	F_1	P	R	F_1	P	R	F_1	F_1
Stanford (Lee et al., 2011)	64.3	65.2	64.7	49.2	56.8	52.7	52.5	46.6	49.4	55.6
Multigraph (Martschat, 2013)	-	-	65.4	-	-	54.4	-	-	50.2	56.7
Unsup. Ranking (Ma et al., 2016)	-	-	67.7	-	-	55.9	-	-	51.8	58.4
c2f-coref	65.7	68.0	66.9	50.9	59.4	54.8	52.9	49.1	50.9	57.5
BERT-base + c2f-coref	66.8	69.2	68.0	51.5	60.6	55.7	53.1	50.3	51.7	58.5
SpanBERT-base + c2f-coref	67.6	68.5	68.1	53.1	60.1	56.4	54.8	50.4	52.5	59.0
BERT-large + c2f-coref	67.2	69.7	68.5	52.3	61.2	56.4	54.0	51.0	52.5	59.1
SpanBERT-large + c2f-coref	67.4	69.8	**68.6**	52.4	61.8	**56.7**	54.1	51.4	**52.7**	**59.3**

Table 1: Results on the test set of the English CoNLL-2012 shared task[3]. The c2f-coref models were trained via weak supervision. Scores for Multigraph and the Unsupervised Ranking model are reported in Ma et al. (2016).

2019), while using their original, respective hyperparameters. We use the implementation of Stanford's system provided with the Stanford CoreNLP suite (Manning et al., 2014). Further training details are provided in Appendix B.

We report precision, recall, and F_1 for the standard MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_{\phi_4}$ (Luo, 2005) metrics. We use the CoNLL F_1 score (average F_1 of the three metrics) as the main evaluation measure, which is common practice in coreference[3].

3.2 Results on OntoNotes

Table 1 shows that the c2f-coref model trained with noisy supervision is able to produce a gain over Stanford's system. The incremental improvement produced by the pre-trained language models highlights the importance of the representation of spans for this task, and suggests that the end-to-end model learns how to effectively exploit it from the noisy supervision. The version of the c2f-coref model augmented with SpanBERT-large achieves 59.3 CoNLL F_1, improving on the Unsupervised Ranking model (Ma et al., 2016) by 0.9 F_1. In contrast with what was observed in the supervised realm (Joshi et al., 2019), the score increase produced by BERT-base over ELMo (+1.0 F_1) is larger than the gain yielded by the large versions of BERT and SpanBERT over their base counterparts (+0.6 and +0.3 F_1, respectively). This might be explained as an effect of the weak supervision, which is likely to reduce the marginal improvement produced by an increase in model complexity.

[3]The metrics are computed using the most recent version of the official CoNLL scorer (Pradhan et al., 2014)

[3]We observed a small discrepancy between the results relative to Stanford's system reported by Ma et al. (2016) and the ones we obtained (~0.2 F_1). Here we report the scores we produce, which are the higher ones.

	Dataset	MUC	B^3	$CEAF_{\phi_4}$	CoNLL
Stanford	PC	59.7	49.7	45.2	51.5
SB-B + c2f	PC	62.0	52.3	47.6	**54.0**
Stanford	LB	65.8	41.6	26.8	44.7
SB-B + c2f	LB	71.4	46.5	31.2	**49.7**

Table 2: Comparison between Stanford's system and the c2f-coref model based on SpanBERT-base (SB-B) on PreCo (PC) and Litbank (LB). Results are expressed in F_1 score.

3.3 Results on PreCo and Litbank

An important feature of PreCo and Litbank is that they contain annotations for singleton mentions, unlike OntoNotes. However, both Stanford's system and the c2f-coref model present a recall-oriented mention detection strategy, which tends to overestimate the number of proposed mentions, as singletons typically would be filtered out from the response. Moreover, the training process of the c2f-coref model does not take singleton mentions into account. For this reasons, we adapt the evaluation on Litbank and PreCo to the OntoNotes guidelines, which assert that predicted singleton mentions should be ignored and non-coreferent spans should be removed from the response. Table 2 shows performance gains consistent with the results on OntoNotes, with the weakly-supervised c2f-coref model improving by 2.5 and 4 CoNLL F_1 on PreCo and Litbank, respectively.

4 Analysis

Performance on Different Types of Coreference
We investigate the capabilities of the weakly supervised end-to-end model in identifying the different kinds of coreference links given by the combination of three mention categories: proper, nominal, and pronominal. We study the performance of the c2f-

Link Type	Stanford	SB-L + c2f	Δ (%)
Nominal - Pronominal	35.7	38.9	+9.0
Nominal - Nominal	54.1	58.6	+8.3
Nominal - Proper	15.1	17.1	+13.2
Pronominal - Proper	60.2	60.4	+0.3
Pronominal - Pronominal	70.9	73.1	+3.1
Proper - Proper	80.8	82.8	+3.5

Table 3: Performance (F_1 scores) on CoNLL-2012 development set in terms of identification of coreference links between different kinds of mentions.

Doc Length	# of Docs	Stanford	SB-L + c2f	Δ (%)
0 - 64	17	52.1	49.6	-4.8
64 - 128	39	57.2	58.6	+2.4
128 - 256	74	56.2	60.9	+8.4
256 - 512	76	58.9	62.3	+5.8
512 - 768	73	56.5	59.6	+5.5
768 - 1152	52	53.3	56.3	+5.6
1152+	12	47.0	50.7	+7.9

Table 4: Average CoNLL F_1 on the OntoNotes development split for sets of documents with different lengths (expressed as number of tokens).

coref model based on SpanBERT-large in comparison to Stanford's system. The results are illustrated in Table 3. We observe a global improvement in all the considered types of links, with the most significant gains from links involving nominal mentions. This improvement is coherent with the observations of Durrett and Klein (2013): coreference decisions involving nominal mentions usually require richer semantic inference, which in our setting is provided by the contextualized span representations

Impact of Document Length We compare the c2f-coref model to Stanford's system on documents of different lengths. As reported in Table 4, Stanford's resolver performs better than the span-ranking system on particularly short documents. However, for all groups of documents longer than 64 tokens, we observe a consistent improvement provided by the c2f-coref model. This could be explained by the contextualized span representations, which were shown to be more informative when larger context is available (Beltagy et al., 2020).

Using Different Linguistic Priors We study how the performance of our approach is impacted as we vary the complexity of the linguistic rules used for the weak supervision. We do this by training the c2f-coref model on the noisy labels obtained using three different implementations of Stanford's system: (1) `1-sieve`, which considers only the Exact String Match rule; (2) `3-sieve`, which consists of the three most effective sieves: Exact String Match, Strict Head Match, and the Pronominal Coreference sieve; and (3) `complete`, which implements all ten sieves. Results in Table 5 show that the improvement provided by the end-to-end model increases as the noisy signal for the training becomes more accurate, suggesting that better supervision helps the model benefit from the knowledge-rich span representations.

Rule Implementation	Stanford	SB-B + c2f	Δ (%)
`1-sieve`	27.9	27.6	-1.1
`3-sieve`	53.5	56.2	+5.0
`complete`	57.0	60.0	+5.3

Table 5: CoNLL F_1 scores on the OntoNotes development set using different combinations of sieves.

Directly facing [***him***]$_1$ *was* [***the box of old***]$_2$ *Mrs. Manson Mingott, whose monstrous obesity had long since made* [***it***]$_2$ *impossible for* [***her***]$_3$ *to attend the Opera...*
Directly facing [***him***]$_1$ *was the box of* [***old Mrs. Manson Mingott***]$_2$*, whose monstrous obesity had long since made it impossible for* [***her***]$_2$ *to attend the Opera...*

Table 6: Example predictions by Stanford's system (upper row) and c2f-coref (lower row) on Litbank. $[\cdot]_x$ represents a mention assigned to cluster x.

Qualitative Analysis In order to better illustrate how the end-to-end system profits from modeling choices unavailable to Stanford's resolver (e.g., contextualized representations), in Table 6 we provide instances of coreference clusters predicted by the two models. The c2f-coref model, unlike Stanford's system, correctly identifies the valid mention *Mrs. Manson Mingott*, links it to the appropriate pronoun (*her*), and correctly neglects the expletive pronoun *it*. This is perhaps because pre-trained models are known to strongly encode syntax (Goldberg, 2019). We present additional examples of predicted chains, along with additional analyses on the impact of the amount of training data, in Appendices C and D, respectively.

5 Conclusion

We presented an approach for coreference resolution that, while being simple, effectively leverages the end-to-end span-ranking model in settings where labeled data is unavailable. Experimental results highlight the efficacy of the weak supervision that the method is based upon, and showed performance gains over previous unsupervised systems.

6 Ethical Considerations

Since our approach is unsupervised and based on the coreference signal produced by Stanford's deterministic coreference system (Lee et al., 2011, 2013), it is prone to echoing biases present in the linguistic rules embodied by Stanford's resolver. Moreover, as most coreference resolvers, the approach we presented is not designed for a particular use case, but it is rather expected to be employed within more complex NLP systems. Specific domains in which these systems are applied (e.g., biomedical data, legal documents) might reveal potential fairness shortcomings in the underlying Stanford's sieve-based system. Depending on the setting of application (e.g., voice assistants or search engines), these possible defects could produce undesirable outcomes. For instance, wrongly classifying two people as the same person is possible to affect information extraction results (e.g., search engines). Further studies on alternative domains are needed to assess these aspects.

Contextual word embedding models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and SpanBERT (Joshi et al., 2020) are pretrained with self-supervised procedures on large portions of unlabeled text. These models are optimized to capture statistical dependencies and might retain and amplify prejudices and stereotypes present in the training data (Kurita et al., 2019). Since the method we propose relies on such pretrained models, it inevitably inherits possible biases that might affect its fairness.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium. 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*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

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: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48, Jeju Island, Korea. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

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

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*, pages 57–60, New York City, USA. Association for Computational Linguistics.

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

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy. Association for Computational Linguistics.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. *arXiv preprint arXiv:2101.00434*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

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.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA. Association for Computational Linguistics.

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.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Xuezhe Ma, Zhengzhong Liu, and Eduard Hovy. 2016. Unsupervised ranking model for entity coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1012–1018, San Diego, California. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Sebastian Martschat. 2013. Multigraph clustering for unsupervised coreference resolution. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 81–88, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, 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.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), pages 8519–8526, Online. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Alexander Wallin and Pierre Nugues. 2017. Coreference resolution for Swedish and German using distant supervision. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 46–55, Gothenburg, Sweden. Association for Computational Linguistics.

Sam Wiseman, Alexander M. Rush, Stuart 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 (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

A Stanford's System

The coreference method proposed by Stanford University at the CoNLL 2011 shared task (Pradhan et al., 2011) is based on a succession of ten independent coreference models (or *sieves*), applied from highest to lowest precision. Here we report a short description of the three most effective sieves, according to Lee et al. (2013).

Exact String Match: links two mentions only if they consist of the exact same text string;

Strict Head Match: implements multiple constraints that must all be matched in order to yield a link. First, the mention head word matches any head word of mentions in the antecedent cluster. Then, all the non-stop words[4] in the cluster of the current mention to be solved are included in the set of non-stop words of the antecedent entity cluster. Moreover, the mention's modifiers (e.g., possessive and personal pronouns) must be all included in the modifiers of the antecedent candidate. Eventually, the two mentions cannot be in an i-within-i construct, (i.e., one must not be a child NP in the other's NP constituent);

Pronominal Coreference Sieve: links pronouns to their compatible antecedents enforcing agreement constraints on a set of attributes, such as gender, number, and animacy.

B Implementation and Training Details

As in previous unsupervised work (Ma et al., 2016), we use the version of the OntoNotes corpus in which the supplementary layers of annotation (e.g., parse trees) were provided automatically using off-the-shelf tools. Using Stanford's system, we obtained the noisy labels for the training and development sets of the CoNLL-2012 shared task data (2802 and 343 documents, respectively), for the PreCo training split (36620 documents), and for Litbank (100 documents). As common practice (Toshniwal et al., 2020), on Litbank we perform 10-fold cross-validation, using sets of 80/10/10 documents for train/development/test.

We trained the models using a batch size of 1 document. On the OntoNotes corpus, the ELMo-based c2f-coref model is trained for a maximum of 150 epochs and the BERT and SpanBERT-based models for 20 epochs. On PreCo and Litbank, the SpanBERT-based c2f-coref model is trained for a maximum of 2 and 400 epochs, respectively. During training, BERT and SpanBERT are fine-tuned. The validation sets used to monitor the training are the development set of OntoNotes and Litbank and a held-out portion of 500 documents from the PreCo corpus. For all datasets, the validation metrics were computed with respect to the Stanford's system-produced noisy labels (i.e., no gold coreference information was used in this process).

We keep the hyperparameter configurations as in Lee et al. (2018) and in Joshi et al. (2020). In particular, for each version of BERT and SpanBERT, we use the combination of `max_segment_len` and learning rates illustrated in table 8.

Training the c2f-coref model based on ELMo, BERT-base and SpanBERT-base took ~6 hours on a 24GB Nvidia TITAN RTX, while the training of the models based on the large versions of BERT and SpanBERT required ~12 hours on a 32GB Nvidia Tesla V100.

	CoNLL F_1
Stanford	57.0
c2f-coref	58.3
BERT-base + c2f-coref	59.1
SpanBERT-base + c2f-coref	60.0
BERT-large + c2f-coref	**60.1**
SpanBERT-large + c2f-coref	**60.1**

Table 7: CoNLL F_1 scores computed on the development set of the CoNLL-2012 shared task.

C Qualitative Examples

Table 9 displays additional examples of coreference chain predictions. In the first example, the weakly-supervised c2f-coref model shows an improved response in terms of both mention identification and cluster assignment, correctly establishing the chains relative to *Alice* and *book*. In example 2, Stanford's system incorrectly links the pronoun *her* to *Mother*, while the neural model rightly associates it with the speaker (*Beth*). Similar improvements are illustrated in sentences 3 and 4. Finally, we report an example of an error propagated from the noisy supervision (sentence 5). Note that singleton mentions were removed from the response cluster, and the mentions that appear as singletons in the reported examples are predicted as coreferent

[4] Stop words are, for instance, *there*, *ltd.*, *etc.*, *'s*.

Model	max_segment_len	bert_learning_rate	task_learning_rate
BERT-base + c2f-coref	128	10^{-5}	$2 \cdot 10^{-4}$
SpanBERT-base + c2f-coref	384	$2 \cdot 10^{-5}$	10^{-4}
BERT-large + c2f-coref	384	10^{-5}	$2 \cdot 10^{-4}$
SpanBERT-large + c2f-coref	512	10^{-5}	$3 \cdot 10^{-4}$

Table 8: Parameters used for the BERT/SpanBERT-based cef-coref models.

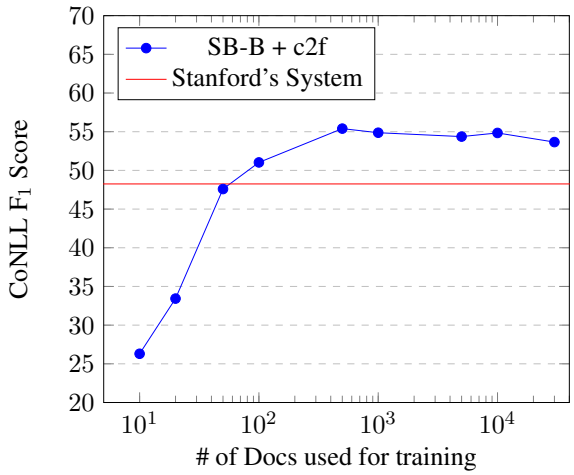

Figure 1: Performance on a held-out set of 1000 PreCo documents using the c2f-coref model as we vary the number of training documents.

to mentions present in other portions of the text.

D Varying the Amount of Training Data

We assess the performance of the model on PreCo when the training is carried out on subsets of different sizes (Fig. 1). We observe that the c2f-coref model requires only 100 weakly-annotated documents to outperform Stanford's system, indicating that the noisy signal is quickly incorporated by the model. Using more than 1000 documents does not seem to boost the score further. We suspect that this behavior might be caused by the homogeneity and the small vocabulary size of the documents of the PreCo dataset.

E Results on the OntoNotes Development Set

We additionally report in Table 7 the results obtained on the development set of the OntoNotes corpus for the five c2f-models.

1	[***CHAPTER I. Down*** [***the Rabbit-Hole Alice***]$_2$]$_1$ *was beginning to get very tired of sitting by* [[***her***]$_2$ ***sister***]$_3$ *on the bank, and of having nothing to do: once or twice* [***she***]$_2$ *had peeped into the book* [[***her***]$_2$ ***sister***]$_3$ *was reading, but* [***it***]$_1$ *had* [***no pictures or conversations in*** [***it***]$_1$]$_4$*, 'and what is the use of a book,' thought Alice 'without* [***pictures or conversations***]$_4$*?'*
	CHAPTER [***I.***]$_1$ *Down the Rabbit-Hole* [***Alice***]$_2$ *was beginning to get very tired of sitting by* [[***her***]$_2$ ***sister***]$_3$ *on the bank, and of having nothing to do: once or twice* [***she***]$_2$ *had peeped into the* [***book***]$_4$ [[***her***]$_2$ ***sister***]$_3$ *was reading, but* [***it***]$_4$ *had no pictures or conversations in* [***it***]$_4$*, 'and what is the use of a book,' thought* [***Alice***]$_2$ *'without pictures or conversations?'*
2	*"*[***We***]$_1$*'ve got* [***Father***]$_2$ *and* [***Mother***]$_3$*, and each other," said* [***Beth***]$_4$ *contentedly from* [***her***]$_3$ *corner.*
	"[***We***]$_1$*'ve got* [***Father***]$_2$ *and* [***Mother***]$_3$*, and each other," said* [***Beth***]$_4$ *contentedly from* [***her***]$_4$ *corner.*
3	*At* [***most terrestrial men***]$_1$ *fancied there might be other men upon* [***Mars***]$_2$*, perhaps inferior to* [***themselves***]$_3$ *and ready to welcome a missionary enterprise.*
	At [***most terrestrial men***]$_1$ *fancied there might be other men upon* [***Mars***]$_2$*, perhaps inferior to* [***themselves***]$_1$ *and ready to welcome a missionary enterprise.*
4	*I persuaded* [***two***]$_1$ *young neighbors to stop playing basketball and to help us get the tree into the house and set* [***it***]$_1$ *correctly in the stand.*
	I persuaded two young neighbors to stop playing basketball and to help us get [***the tree***]$_1$ *into the house and set* [***it***]$_1$ *correctly in the stand.*
5	*To prevent* [***this***]$_1$*, humans on* [***Mars***]$_2$ *have to wear special shoes to make* [***themselves***]$_1$ *heavier.*
	To prevent [***this***]$_1$*, humans on* [***Mars***]$_2$ *have to wear special shoes to make* [***themselves***]$_1$ *heavier.*

Table 9: Example predictions by Stanford's system (upper sub-row) and c2f-coref (lower sub-row) on Litbank (examples 1-3) and PreCo Dev (examples 4 and 5). $[\cdot]_x$ represents a mention assigned to cluster x.