Evaluation of Named Entity Coreference

Oshin Agarwal [*]
University of Pennsylvania
oagarwal@seas.upenn.edu

Sanjay Subramanian [*]
University of Pennsylvania
subs@seas.upenn.edu

Ani Nenkova
University of Pennsylvania
nenkova@seas.upenn.edu

Dan Roth
University of Pennsylvania
danroth@seas.upenn.edu

Abstract

In many NLP applications like search and information extraction for named entities, it is necessary to find all the mentions of a named entity, some of which appear as pronouns (she, his, etc.) or nominals (the professor, the German chancellor, etc.). It is therefore important that coreference resolution systems are able to link these different types of mentions to the correct entity name. We evaluate state-of-the-art coreference resolution systems for the task of resolving all mentions to named entities. Our analysis reveals that standard coreference metrics do not reflect adequately the requirements in this task: they do not penalize systems for not identifying any mentions by name to an entity and they reward systems even if systems find correctly mentions to the same entity but fail to link these to a proper name (she–the student–no name). We introduce new metrics for evaluating named entity coreference that address these discrepancies and show that for the comparisons of competitive systems, standard coreference evaluations could give misleading results for this task. We are, however, able to confirm that the state-of-the art system according to traditional evaluations also performs vastly better than other systems on the named entity coreference task.

1 Introduction

Coreference resolution is the task of identifying all expressions in text that refer to the same entity. In this paper we set out to provide an in-depth analysis of the task specifically for named entities: finding all references—either by name, pronoun or nominal—to a named entity in the text.

Many language technology tasks focus on entities and our work is oriented towards practical uses of the results of coreference resolution in downstream tasks. Named entities are often targets for information extraction (Ji and Grishman, 2011), biography summarization (Zhou et al., 2004) and knowledge base completion tasks (West et al., 2014). More relevant information can be extracted for these tasks if we also know which pronouns and nominals refer to the entity. Similarly, creation of proper noun ontologies (Mann, 2002) can use patterns other than (proper noun–common noun) if other references to the entity are known.

Recent work (Webster et al., 2018) has shown that standard coreference datasets are biased and high performance on these need not mean high performance in downstream tasks. We argue that the standard coreference metrics are not suitable either from the perspective of downstream applications. Since applications require information about entities and entities are usually identified by their names, the evaluation metrics should focus on the resolution of mentions to the correct name. If all the pronouns referring to an entity are resolved correctly to each other but are not linked to any name or are linked to a wrong name, the results would not be useful for downstream tasks. Standard coreference metrics do not incorporate these aspects and hence give high performance for results unsuitable for further use. We also show that the existing metrics are not sensitive to finding any mention to an entity at all. They give higher performance for systems that do not find a large number of entities but do good coreference resolution on the subset of entities they find.

This problem of coreference chains without any named mentions being unsuitable has previously been discussed in (Chen and Ng, 2013). The authors argued that a name is more informative than a nominal, which is more informative than a pronoun so they assign different weights to co-reference links (mention-antecedent pairs) in a chain depending on the type of mentions the link contains. They assign a higher weight to

[*] equal contribution

Proceedings of the 2nd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2019), pages 1–7,
Minneapolis, USA, June 7, 2019. ©2019 Association for Computational Linguistics

a link having a name than one that doesn't and also higher weight to a link having a nominal than a link that contains just pronouns. Similarly, (Martschat and Strube, 2014) perform an error analysis for co-reference by choosing an antecedent that is a name or a nominal in this order because they are more informative than a pronoun. However, we argue that we should view the coreference chains as a whole instead of individual links when evaluating systems for downstream application. If a chain contains even one named mention, it should be sufficient for using it in applications and we need not consider the mention type in each link within the chain.

We introduce metrics focused on Named Entity Coreference (NEC) which separate the identification of entities and resolution of different mention types, thus tackling the above issue and transparently tracking areas of system improvement.

2 Coreference Evaluation

Shared tasks on coreference (at CoNLL-2011 and 2012 (Pradhan et al., 2014)) use the average of three F1 scores as their official evaluation: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and CEAFE (Luo, 2005). Prior work (Moosavi and Strube, 2016) discussed shortcoming of these metrics and introduced the improved link entity aware (LEA) score. Below we describe each score in the context of downstream tasks. Let K be the set of key (gold) clusters, and let R be the set of response clusters.

MUC The recall for an entity is the minimum number of links that would have to be added in the predicted clusters containing any mention of this entity, to make them connected and part of the same cluster. Precision is computed by reversing the role of gold and predicted clusters.

$$\text{Recall} = \frac{\sum_{k_i \in K} |k_i| - |p(k_i)|}{\sum_{k_i \in K}(|k_i| - 1)}$$

where $p(k_i)$ is the partition of k_i generated by intersecting k_i with the response entities.

Gold: {JohnDoe, he_1, he_2, he_3} {RichardRoe, he_4, he_5}
Solution 1: {JohnDoe, he_1, he_2} {RichardRoe, he_4}
Solution 2: $\{he_1, he_2, he_3\}\{he_4, he_5\}$

Table 1: Hypothetical Solution 2 has no practical value.

B-cubed B^3 works on the mention level. It iterates over all gold-standard mentions of an entity, averaging the recall of its gold cluster in its predicted cluster. It computes precision by reversing the role of gold and predicted clusters.

$$\text{Recall} = \frac{\sum_{k_i \in K} \sum_{r_j \in R} \frac{|k_i \cap r_j|^2}{|k_i|}}{\sum_{k_i \in K} |k_i|}$$

CEAF CEAF first maps each gold cluster to a predicted cluster. It then computes recall as the number of similar mentions shared by the gold and predicted clusters divided by the number of mentions in the gold cluster. Precision is equal to the number of similar mentions shared by the gold and predicted, divided by the number of mentions in the predicted cluster. Numbers are reported either per mention (CEAFm), or per entity (CEAFe).

$$\text{Recall} = \frac{\sum_{k_i \in K^*} \phi(k_i, g^*(k_i))}{\sum_{k_i \in K} \phi(k_i, k_i)}$$

where K^* is the set of key entities in the optimal one-to-one mapping and $\phi(\cdot, \cdot)$ is a similarity measure for a pair of entities. In CEAFm, $\phi(k_i, r_j) = |k_i \cap r_j|$, and in CEAFe, $\phi(k_i, r_j) = \frac{2|k_i \cap r_j|}{|k_i| + |r_j|}$.

LEA Recall is computed as the fraction of correctly resolved links between mentions. Results for each entity are weighted by its number of mentions, so that resolving correctly an entity with more mentions contributes more to the overall score. Precision is computed by reversing the role of gold and predicted clusters.

$$\text{Recall} = \frac{\sum_{k_i \in K} \left[|k_i| \times \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)} \right]}{\sum_{k_i \in K} |k_i|}$$

where for any set S, $link(S)$ denotes the number of links between elements of S (so $link(S) = |S| \cdot (|S| - 1)/2$).

	Solution 1			**Solution 2**		
	R	**P**	**F1**	**R**	**P**	**F1**
MUC	0.60	1	0.74	0.60	1	0.74
B-cub	0.51	1	0.67	0.51	1	0.67
CEAFm	0.71	1	0.83	0.71	1	0.83
CEAFe	0.82	0.82	0.82	0.82	0.82	0.82
LEA	0.42	1	0.60	0.42	1	0.60
NEC	0.71	1	0.83	0	1	0

Table 2: Evaluation of the hypothetical examples in Table 1. NEC is the new metric introduced in Section 3.

The goal of NEC is to link all mentions referring to a named entity to the correct name. Consider the example in Table 1. There are two entities, each with one named mention and a few pronouns. Both solutions find the same number of correct mentions pairs. However, solution 1 has a named mention in each cluster but solution 2 has only pronouns. Standard evaluations have the same values for both solutions (see Table 2) because they do not consider the types of mentions.

3 NEC Evaluation Metrics

The above example highlights the potential deficiencies of standard coreference evaluations when applied to NEC. Here we introduce a set of task-specific criteria for the evaluation of NEC.[1]

3.1 NEC F1

In the gold-standard, all mentions to named entities are grouped into chains. We wish to find a chain corresponding to each entity in the system output also. To map chains between the gold-standard and the system output, we select for each gold-standard chain, the predicted chain that has the highest F1 score with respect to its mentions. The NEC F1 score is the average of these highest per entity scores.[2]

To compute the intersection between a gold-standard and a system chain, we first augment each gold-standard chain with a list of all variations of the entity's name. We rely on the gold-standard named entity annotation and intersect this with the membership in a coreference chain. This provides lists of the full name, last name, occasionally nicknames, i.e. {*Frank Curzio, Francis X. Curzio, Curzio*}, {*Dwayne Dog Chapman, Dog Chapman, Chapman*}. We consider a predicted chain to be a candidate match for a gold chain only if it contains at least one of the name variants. We do not use exact mention match to find candidate chains as the presence of the name can indicate which entity the cluster is about. If the gold mention is 'Mr Joe from Boston' and the system finds 'Mr Joe', we still consider the chain containing this mention to be a candidate chain as the name can be determined and other mentions may have been resolved correctly.

For each named entity $k_i \in K$, let N_i be the set of response mentions that contain the full name of k_i. For a key named entity k_i and a response entity r_j, the precision is defined to be $p(k_i, r_j) = \frac{|r_j \cap k_i|}{|r_j|}$ and the recall is defined to be $r(k_i, r_j) = \frac{|r_j \cap k_i|}{|k_i|}$. The $F1$ for this pair of key entity and response entity is then given by $f(k_i, r_j) = \frac{2p(k_i,r_j)r(k_i,r_j)}{p(k_i,r_j)+r(k_i,r_j)} = \frac{2|r_j \cap k_i|}{|r_j|+|k_i|}$. Then F1 for the key named entity k_i is

$$\mathrm{F1}_i = \max_{r_j \in R : r_j \cap N_i \neq \emptyset} f(k_i, r_j)$$

We use an exact span matching between gold and predicted mentions to calculate F1 to be consistent with the existing scorers.

If a gold-standard chain does not get paired with any system chain, the F1 for that chain is taken to be zero. We find the overall F1 of the system as the average of the F1 for each gold chain, $\frac{1}{|K|} \sum_{k_i \in K} \mathrm{F1}_i$.

3.2 Entity not Found

The NEC F1 gives a sense of overall performance but mixes true purity of the system-discovered entities and the ability to discover entities at all. "Entity not found" is the error when no NEC system output overlaps with a gold standard chain. These contribute a score of 0 for the average F1.[3]

3.3 Pronoun Resolution Accuracy

We also track the NEC F1 when only mentions of given syntactic type are preserved in the chain—name, pronoun and nominal. Of special interest is to track performance when resolving pronouns. Many of the errors on pronouns arise due to the need for common-sense knowledge and reasoning.

3.4 Over-Splitting/Combination of Entities

We tracked the over-splitting (systems produce multiple clusters for the same name) and the over-combination of entities as well (placing mentions to different named entities in the same cluster. This error usually occurs when different people have the same last name but also occasionally when the names are completely different but the roles of the people are similar. However, overall

[1] See supplementary material for examples of the errors.

[2] Although the task appears similar to Entity Linking (EL) (Mihalcea and Csomai, 2007; Ratinov et al., 2011), it does not involve linking an entity to a knowledge base (KB). Not all entities even need to be in a KB. Also, EL typically focuses on names and other nouns whereas coreference includes pronouns as well.

[3] We consider only chains containing a named mention. Chains that do not contain any named mention are filtered out. More details on filtering in section 4.

	PER			ORG			GPE		
	Chains not found	NEC F1	Coref F1	Chains not found	NEC F1	Coref F1	Chains not found	NEC F1	Coref F1
(Raghunathan et al., 2010)	16%	0.55	0.50	34%	0.42	0.41	14%	0.67	0.56
(Clark and Manning, 2015)	36%	0.46	0.56	40%	0.39	0.46	21%	0.61	0.61
(Clark and Manning, 2016a,b)	21%	0.61	0.67	29%	0.50	0.52	17%	0.68	0.65
(Lee et al., 2017)	28%	0.58	0.69	26%	0.58	0.56	12%	0.76	0.68
(Lee et al., 2018)	7.5%	0.80	0.77	15%	0.69	0.61	8%	0.81	0.69

Table 3: Performance of systems. Chains not found and NEC F1 refer to the new named entity focused metrics. Coref F1 refers to the evaluation combining MUC, B^3 and CEAFE, on test data.

	PER			ORG			GPE		
	Name	Pronoun	Nominal	Name	Pronoun	Nominal	Name	Pronoun	Nominal
(Raghunathan et al., 2010)	0.55	0.45	0.23	0.47	0.35	0.11	0.73	0.44	0.19
(Clark and Manning, 2015)	0.50	0.34	0.10	0.46	0.34	0.15	0.65	0.57	0.22
(Clark and Manning, 2016a,b)	0.66	0.49	0.15	0.54	0.47	0.33	0.72	0.59	0.41
(Lee et al., 2017)	0.64	0.41	0.15	0.65	0.48	0.39	0.80	0.70	0.47
(Lee et al., 2018)	0.85	0.58	0.26	0.76	0.64	0.47	0.85	0.77	0.51

Table 4: NEC F1 by type of mention. The errors on names are high, though it is possible to resolve these with NER and string matching or similarity. Pronoun errors are high as expected.

such errors were quite small and similar for all systems and have thus not been included in the later tables with results.

4 Evaluation of Systems

We make use of the relevant part of OntoNotes coreference corpus (Pradhan et al., 2007) and gold-standard annotations for named entities on the same data to quantify the patterns in coreference of different named entity types (see the table in the supplementary material) and to evaluate systems on the newswire, broadcast news and magazine documents for PER, ORG and GPE entities.

Patterns in Coreference Named people, organizations and locations make up 38% of all coreference clusters in OntoNotes (Pradhan et al., 2007), yet 54% of all mentions that require coreference resolution are mentions of these types. All named entities are on average much less likely to be singletons than a typical entity, mentioned only once in the text and not requiring coreference resolution (De Marneffe et al., 2015). People, organizations and locations are most likely to be mentioned repeatedly: 68% of people, 51% of organizations and 52% of locations named in text have at least one other coreferent mention to them.

Named entities have a large portion of references that are not by name. Nominals account for less than 5% of the mentions in all genres for PER, while the remaining mentions are split almost equally between names and pronouns. For ORG, roughly half of the mentions are named, the remaining are equally split between pronouns and nominals. For GPE, roughly 70% of the mentions are named and others are mostly pronouns.

Systems We evaluate the Stanford coreference system, with its deterministic (Raghunathan et al., 2010; Lee et al., 2011; Recasens et al., 2013), statistical (Clark and Manning, 2015) and neural (Clark and Manning, 2016a,b) versions, and the neural end-to-end systems of (Lee et al., 2017) and (Lee et al., 2018) on traditional and NEC metrics.

These general coreference systems find coreferring expressions of any type and produce coreference chains for all mentioned entities. In NEC, the goal is to find all mentions to *an entity* that has been *referred to by name* at least once in the document. The output of off-the-shelf coreference systems has to be filtered to keep only chains that contain at least one mention noun phrase with a syntactic head that is a entity's name.[4] For our evaluation, we use the spaCy dependency parsing system (Honnibal and Johnson, 2015) to detect whether a name is the head of a mention, by checking that no other word in the mention is an ancestor of the name in the dependency parse tree. In evaluation, we use gold NER tags to determine if the head is a name. Note that the dependency parsing and gold NER are not given to the systems but are used to process their output.

Many system NEC chains did not have any

[4] Less strict filtering, such as the presence of an appropriate pronoun could also indicate that it a specific type of entity. For NEC, we insist on having at least one named mention.

named mentions. (Lee et al., 2017) does not have a named mention in about 30% of the coreference chains on PER that do contain a personal or possessive third person pronoun. This number is about 20% for the CoreNLP neural system.

Table 3 shows the standard and NEC F1 on all the systems. For PER, there are three notable leaps of improvement according to the standard coref evaluation: between the statistical and rule-based CoreNLP systems, between their statistical and neural systems and between the two versions of the AllenNLP systems. Some of these improvements contradict actual performance on NEC, notably for the difference between the rule-based and statistical systems. The other two improvements in Coref F1 translate to improvements in NEC metrics. The difference between the statistical and rule-based system is also falsely reflected in standard F1 for ORG and ORG entities. As expected, (Lee et al., 2018) outperforms all the systems, with (Lee et al., 2017) as a close second. Both perform much better than (Raghunathan et al., 2010) and (Clark and Manning, 2015). (Clark and Manning, 2016a,b) does slightly better than (Lee et al., 2017) on PER entities. Notably, (Lee et al., 2018) misses less than 10% of the chains for all entity types compared to 20-40% by other systems.

Note that the performance varies considerably across entity types. A top NER system such as (Ratinov and Roth, 2009) that focus on PER, ORG and GPE does not find a single named entity in just 4.67%, 5.7% and 1.1% of chains respectively. However, the percentage of chains not found is much higher. It is possible that the non-named mentions were resolved to each other but not to any names so such chains got filtered out for the NEC task. Future work involves developing coreference systems driven by NER and producing results more suitable for downstream tasks.

We also separate the performance of the systems by mention type. The second panel of Table 3 reveals that (Lee et al., 2018) outperforms all the systems on each mention type for all the three types of entities. Detection of named mentions can be done with high accuracy by named entity recognition systems (Stoyanov et al., 2009) and the matching of names can also be done accurately via string matching (Wacholder et al., 1997; Wick et al., 2009). In spite of this, most systems do not perform well on names. The mistakes on pronouns and nominals are much higher as expected.

While (Lee et al., 2018) gets a better F1 on the standard coreference metrics used as well, it improves on many aspects of performance. It finds more chains and even performs better resolution of each mention type, making it more suitable for downstream tasks.

5 Conclusion

We presented the task of Named Entity Coreference (NEC) and argued that the standard coreference metrics are not suitable for the evaluation of this task. We introduced evaluation metrics that tackle the shortcomings of the standard metrics for the task and track the different errors made by systems. We showed that many off-the-shelf systems do not perform well on these metrics. They output many clusters without a link to any name or a link to the incorrect name, making results unsuitable for downstream applications. Our metrics track different aspects of system performance and help identify such issues.

Acknowledgments

This work was supported in part by Contract HR0011-15-2-0025 with the US Defense Advanced Research Projects Agency (DARPA). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Amit Bagga and Breck Baldwin. 1998. Entity–based cross–document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference*, pages 79–85.

Chen Chen and Vincent Ng. 2013. Linguistically aware coreference evaluation metrics. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1366–1374.

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. 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 1405–1415.

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.

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.

Marie-Catherine De Marneffe, Marta Recasens, and Christopher Potts. 2015. Modeling the lifespan of discourse entities with application to coreference resolution. *J. Artif. Int. Res.*, 52(1):445–475.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1*, HLT '11, pages 1148–1158.

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 15th conference on computational natural language learning: Shared task*, pages 28–34. 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.

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.

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

Gideon S Mann. 2002. Fine–grained proper noun ontologies for question answering. In *Proceedings of the 2002 workshop on Building and using semantic networks–Volume 11*, pages 1–7. Association for Computational Linguistics.

Sebastian Martschat and Michael Strube. 2014. Recall error analysis for coreference resolution. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2070–2081.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge.

Nafise Sadat Moosavi and Michael Strube. 2016. 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)*, pages 632–642.

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 S Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing*, 1(04):405–419.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*, pages 147–155.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 627–633.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state–of–the–art. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore*, pages 656–664.

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*, MUC6 '95, pages 45–52.

Nina Wacholder, Yael Ravin, and Misook Choi. 1997. Disambiguation of proper names in text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pages 202–208.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge base completion via search-based question answering. In *Proceedings of the 23rd international conference on World wide web*, pages 515–526. ACM.

Michael Wick, Aron Culotta, Khashayar Rohanimanesh, and Andrew McCallum. 2009. An entity based model for coreference resolution. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 365–376. SIAM.

Liang Zhou, Miruna Ticrea, and Eduard Hovy. 2004. Multi–document biography summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.