Supervised Within-Document Event Coreference using Information Propagation

Zhengzhong Liu, Jun Araki, Eduard Hovy, Teruko Mitamura
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA
liu@cs.cmu.edu, junaraki@cs.cmu.edu, hovy@cmu.edu, teruko@cs.cmu.edu

Abstract

Event coreference is an important task for full text analysis. However, previous work uses a variety of approaches, sources and evaluation, making the literature confusing and the results incommensurate. We provide a description of the differences to facilitate future research. Second, we present a supervised method for event coreference resolution that uses a rich feature set and propagates information alternatively between events and their arguments, adapting appropriately for each type of argument.

Keywords: event coreference, information extraction, event coreference evaluation

1. Introduction

Coreference resolution, the task of linking surface mentions to their underlying discourse entities, is an important task in natural language processing. Most of the early work focused on coreference of entity mentions. Recently, event coreference has attracted attention on both theoretical and computational aspects. However, most event coreference work is preliminary and applied in quite different circumstances, making comparisons difficult or impossible.

In this paper, we first provide an overview of all the relevant literature to identify the ways each experiment differs from the others. Our claim here is that the comparisons to related work in prior papers are not really appropriate due to these differences. This makes future research difficult. We then present a supervised approach to event coreference, and describe a method for propagating information between events and their arguments that can improve results. In our method, different argument types support different methods of propagation. For these experiments, we annotate and use a corpus of 65 documents in the Intelligence Community (IC) domain that contains a rich set of within-document coreference links (Hovy et al., 2013).

2. Related Work

Table 1 summarizes recent work on event coreference resolution. For the reasons below, only one supervised system (Ahn, 2006) and two unsupervised (Bejan & Harabagiu, 2010; Cybulska & Vossen, 2012) on within-document event coreference are suitable as a basis for ongoing comparison.

2.1. Problem definition:

Different approaches use different definitions of the problem (see Compatible Definition column). However, as discussed in recent linguistic studies (Recasens et al., 2011; Hovy et al., 2013), the existence of different types and degrees of coreference makes it necessary to agree on the definition of cofererence before performance can be compared. The lack of clarity about what coreference should encompass rules out several systems for comparison. OntoNotes created restricted event coreference (Pradhan et al., 2007), coreferring only some nominalized events and some verbs, and not reporting event-specific results. Both Naughton (2009) and Elkhlifi and Faiz (2009) worked on sentence-level coreference, which is closer to the definition of Danlos (2003). However it is unclear when one sentence contains multiple event mentions, and hence these are not comparable to systems that process more specific coreference units.

2.2. Dataset and settings:

Early work by Bagga and Baldwin (1999) conduct experiments only on cross-document coreference. Recent advanced work on event coreference is by Bejan and Harabagiu (2010) and Lee et al. (2012) use the ECB corpus[1](or a refined version[2]) to evaluate performance, which is annotated mainly for cross-document coreference. In this corpus, within-document coreference is only very partially annotated; most difficult coreference instances are not marked.

D1	S1	*Indian naval forces **came to the rescue (E1)** of a merchant vessel under **attack (E3)** by pirates in the Gulf of Ade on Saturday, **capturing (E2)** 23 of the raiders , India **said (E4)**.*
	S5	*Indian commandos boarded the larger pirate boat, **seizing** 12 Somali and 11 Yemeni nationals as well as arms and equipment, the statement **said**.*
D2	S1	*The Indian navy **captured (E2)** 23 piracy suspects who **tried (E5)** to **take over (E3)** a merchant vessel in the Gulf of Aden, between the Horn of Africa and the Arabian Peninsula, Indian officials **said (E4)** .*
	S3	*In addition to the 12 Somali and 11 Yemeni suspects, the Indian navy **seized** two small boats and "a substantial cache of arms and equipment," the military **said** in a statement.*

The examples above are extracted from two documents from the ECB. Dx and Sx denote document id and sentence id respectively. In both documents S1 is annotated once, but not in the rest of the article. In D1, we find in S5 the event mention "**seizing**" which should actually be coreferent with

[1]http://adi.bejan.ro/data/ECB1.0.tar.gz
[2]http://nlp.stanford.edu/pubs/jcoref-corpus.zip

	Gold standard used	Cross/within Document		Compatible definition	Corpus
		Within	Cross		
This paper	Mention head word	✓		✓	IC
Lee et al. (2012)			✓	✓	ECB
Sangeetha and Arock (2012)	ACE mention, arguments and attributes	✓		✓	ACE
Cybulska and Vossen (2012)	Mention head word	✓		✓	IC
McConky et al. (2012)	ACE mention, arguments and attributes	✓		✓	ACE
Li et al. (2011)	Human entity and event mention detection	✓		✓	Unavailable
Bejan and Harabagiu (2010)		✓(ACE, ECB)	✓(ECB)	✓	ECB, ACE
Chen and Ji (2009)	ACE mention, arguments and attributes	✓		✓	ACE
Elkhlifi and Faiz (2009)		✓			Unavailable
Naughton (2009)		✓			IBC, ACE
Pradhan et al. (2007)		✓			OntoNotes
Ahn (2006)	ACE mention, arguments and attributes	✓		✓	ACE
Bagga and Baldwin (1999)			✓		Unavailable

Table 1: Recent computational approaches to event coreference resolution.

"**capturing (E2)**" in D1;S1. In D2;S3, we find a more tricky case: the mention "**seized**", which has semantics similar to "**captured**" but is not coreferent due to different patients. The cross-document case also doesn't seem to compatible with our definition. In ECB, "**attack (E3)**" in D1;S1 is annotated as coreferent with "**take over (E3)**" in D2;S1, which we believe is wrong: at best, the attack is only a part of the attempt to take over the merchant vessel.

Goyal et al. (2013) use a distributional semantic approach on event coreference. However, they didn't adopt a conventional evaluation setting. They draw from the IC corpus an equal number of positive and negative testing examples, which is different from the natural data distribution.

2.3. Gold standard annotations used:

Recent work using the ACE 2005 corpus[3] (Chen & Ji, 2009; Chen et al., 2009; Sangeetha & Arock, 2012; McConky et al., 2012; Ahn, 2006) agrees with our definition of coreference. However, the ACE corpus annotations, in addition to event mentions, also include argument structures, entity ids, and time-stamps. Most coreference systems on the ACE corpus make use of this additional information. This makes them impossible to compare to systems that do not make this simplifying assumption. It also makes results achieved on ACE hard to compare to results on corpora without this additional information. Among these work, only Ahn (2006) reported some results using system generating arguments, we compare our system against it.

2.4. Availability

Li et al. (2011) use a hand-annotated web corpus, which is not publicly available for comparison.

In summary, anyone wanting to work on within-document event coreference has to obtain a corpus that is fully annotated, that does not include additional facilitating information, whose definition of coreference respects the theoretical considerations of partial coreference, and that has other systems freely available for comparison. Meeting these criteria is not easy. The closest work we find is by Cybulska and Vossen (2012) and Bejan and Harabagiu (2010), both adopt unsupervised methods for event coreference. Ahn (2006) also reported results on ACE by swapping gold standard annotations with system results.We compare our system to their results on their corresponding corpus.

[3]http://www.itl.nist.gov/iad/mig/tests/ace/2005/doc/

	Total	Avg.
Event Mention	2678	41.2
Non-elliptical Domain Event Mention	1998	30.7
Reporting Event Mention	669	10.29
Full coreference relations	1253	21.6
Subevent relations (parent-child)	455	8
Membership relations (parent-child)	161	2.9

Table 2: Corpus Statistics

3. Corpus

Our system is trained and evaluated on the IC domain corpus, which annotates several different event relations. Table 2 summarizes the corpus level statistics and the average over documents. In this work, we focus on full coreference relations. The inter-annotator agreement among 3 annotators for full coreference is 0.614 in terms of Fleiss's kappa (Fleiss, 1971). For detailed definition for the corpus, we refer readers to Hovy et al. (2013). To facilitate future research, We also report our system results on the ACE 2005 training dataset, which contains 599 documents.

4. System description

Our system is almost end-to-end, except that we start with a minimal gold standard head word annotations in order to focus on the core coreference problem. This approach is the same as Cybulska and Vossen (2012) and Bejan and Harabagiu (2010)[4].

4.1. Procedure

Similar to Chen et al. (2009), we approach the problem first with a conventional pairwise model:
1) Supervised classification that determines the probability whether two mentions corefer. The classifier used in the experiment is Random Forest (Breiman, 2001), implemented in Weka (Hall et al., 2009).
2) Clustering that processes all the pairwise scores to output the final clusters of pairs.
3) In addition, we added a third step after clustering, information is propagated between event mentions to enrich the original feature set. Typically, the information carried from one event to its coreferent mention is about the participants (agent, patient, etc.). When an event has been enriched by receiving information from another, it may in turn now be linkable to a third event. The system repeats this process until no more information can be propagated. Currently, the propagation includes two parts: 1) if one mention has missing arguments, they will be copied over from the coreferred counterpart; 2) if both arguments are present, information not presented in one will be copied from another.
Similarly, Lee et al. (2012) show that jointly modeling references to events and entities can boost the performance on both. We hold a similar assumption. But by focusing on events and their arguments, we can perform propagation specific to each type of argument, for instance, geographical reasoning as described below.

4.2. Features

In addition to typical lexical and discourse features, we also model an event mention with its surface form and its arguments, including agent, patient[5], and location. We use a rich set of 105 semantic features, described in table 3.

4.2.1. Agent, patient extraction and propagation

We use the semantic parser *Fanse* (Tratz & Hovy, 2011) to annotate the predicate arguments defined in PropBank. For nominal events, we extract agent and patient using heuristics such as finding the token attached to the event mention with specific words (such as "by") and modifiers as agent (e.g., HAMAS in HAMAS's attack). During the propagation step, information not present in one entity can be copied from another.

4.2.2. Location extraction and propagation

In contrast to agent and patient, the propagation of location information employs external information to gain additional power. We use the *Stanford Entity Recognition* (Finkel et al., 2005) engine to identify location mentions. *DBpedia Spotlight* (Mendes et al., 2011) is run to disambiguate location entities. DBpedia (Lehmann et al., 2014) information, such as cities, country, and alternative names, are then injected. When the location is not found in DBpedia, we search the mention string in *Geonames*[6] and use the first result with highest Dice coefficient with the surface string. This world knowledge enriches annotation. For example, we can now match the mention "Istanbul" with the country name "Turkey".

4.3. Clustering

We conduct experiments with two simple clustering methods. The first is a pure transitive closure that links all pairs mentions that the classification engine judges as positive. The second is the Best-Link algorithm of Ng and Cardie (2002), which links each mention to its antecedent with the highest likelihood when the classifier judges as positive.

5. Evaluation

5.1. Evaluation Metrics

Coreference evaluation metrics have been discussed by the community for years. To enable comparison, we report most metrics used by the CoNLL 2012 shared task (Pradhan et al., 2012), including MUC (Chinchor & Sundheim, 1993), B-Cubed (Bagga & Baldwin, 1998), entity-based CEAF (Luo, 2005), and BLANC (Recasens & Hovy, 2011). Pairwise scores are used to provide a direct view on performance.

5.2. Experiments and Results

We split the documents in IC corpus randomly into 40 documents for training and development, and 25 for testing. Parameters such as the probability threshold to determine coreference are tuned on the 40 documents using five-fold cross validation. Optimization is not done separately for each metric. We simply use a universal classifier threshold optimized for pairwise case. During experiment, the propagation step is actually performed for only one iteration, since no further information is propagated. On the ACE corpus, we simply apply the best model configuration from IC corpus and train on 90% of the documents (539) for training and 10% for testing (60).
Table 3 summarizes the overall average results obtained by BestLink on both ACE and IC corpus (BestLink consistently outperforms naively full transitive closure). We also attach three other reported results at the end. Note that these results are not directly comparable: Cybulska and Vossen (2012) and Bejan and Harabagiu (2010) use unsupervised methods, thus their reported results are evaluated on the whole corpus; Ahn (2006) also use a 9:1 train-test split, but the split might be different with ours. A simple comparison shows that our results outperform these systems in all metrics, which is notable because all these metrics are designed to capture the performance from different aspects.
To interpret the results, it should also be noted that because of the existence of large number of singleton clusters, some measures such as B^3 seem to be high even using the most

[4] Although Bejan and Harabagiu (2010) use automatic mention detection to extend the mention set for training, they only use true mentions of the ACE dataset at evaluation time.

[5] Specifically, these are defined as ARG0, ARG1 in PropBank. They could be more-specific variants roles such as *experiencer*, but we prefer a smaller set for simplicity.

[6] http://www.geonames.org/

	Pairwise			MUC			B^3			CEAF-e			BLANC		
IC corpus	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
Discourse + Lexical	32.69	25.11	28.40	41.7	33.58	37.2	79.46	74.06	76.67	66.89	73.95	70.24	59.77	61.2	60.43
+ Syntactic	47.12	35.15	40.26	52.6	47.63	50.0	82.24	81.46	81.85	76.91	80.21	78.53	64.76	68.59	66.42
+ Semantic (no arguments)	51.15	42.22	46.26	54.5	49.1	51.68	82.12	82.08	82.1	74.93	78.31	76.58	65.41	69.98	67.35
+ Arguments	55.96	47.86	51.60	56.87	**55.81**	56.33	83.38	**85.58**	**84.46**	**88.13**	80.73	**80.43**	68.77	**75.21**	71.46
+ Propagation	**59.04**	**48.27**	**53.11**	**68.72**	55.5	**61.44**	**89.28**	79.89	84.33	75.14	**82.9**	78.83	**82.28**	70.77	**75.06**
Cybulska and Vossen (2012)	-	-	-	-	-	-	81.0	71.0	76.0	-	-	-	-	-	-
ACE corpus															
This work	55.86	40.52	**46.97**	53.42	48.75	50.98	89.9	88.86	**89.38**	85.54	87.42	**86.47**	70.88	70.01	70.43
Bejan and Harabagiu (2010)	43.3	47.1	45.1	-	-	-	83.4	84.2	83.8	76.9	76.5	76.7	-	-	-
Ahn (2006)	-	-	43.3	-	-	-	-	-	-	-	-	-	-	-	-

Table 3: Evaluation results and comparisons

naive feature set. By looking at the pairwise performance, however, we see that current best F-score is only about 50%. There are still many challenges in event coreference.

6. Discussion

The evaluation results show that almost all types of features help to improve the performance over all metrics rather consistently. However, preliminary error analysis shows that some events are still clustered incorrectly even when arguments match. We argue that limitations in argument extraction and entity coreference prevent these features from contributing directly to correct coreference decisions. On the other hand, the results of propagation show that new information helps to find more links but inevitably comes with a drop in precision. We consider that modeling event and arguments holistically like Lee et al. (2012) would help guide the propagation. By inspecting the data, we hypothesize that the main benefits brought by the propagation scheme is to match arguments of two coreferent events. If the arguments are nominal events, they will be then marked as coreferent due to the feature "Event as Entity" (See Semantic features in table 4). In the following example, if the two event mentions "planning" are marked as coreference, then the corresponding argument "attack" will be also marked as coreference.

A member of the Islamic militant movement HAMAS suspected of ***planning*** *a suicide* ***attack*** *against Israel surrendered to Palestinan police here after a six-hour shootout on Friday.*

HAMAS's military wing, was on the run from both Palestinian and Israeli police for ***planning*** *anti-Israeli* ***attacks****.*

This hypothesis is also in line with our observation that propagation can only be performed for one round, because the nominal event themselves are unlikely to have other nominal events as arguments. Such interactions between event mentions also remind us that conference can be possibly improved by other types of event relations, such as subevent relations.
Furthermore, the system tends to merge clusters where the event mention head words are the same because the head word feature receives a high weight in the model, even when this is not appropriate. More work should be performed on disambiguating such difficult cases.

7. Conclusion and Future Work

In this paper we first describe why most previous work on coreference is not directly comparable to one another, for a variety of reasons. In particular, reports of high coreference performance on one corpus do not really transfer over to other corpora or other definitions of coreference. Event coreference is not a solved problem.
We then present a simple supervised pairwise event coreference system. We show that rich linguistic features, especially event arguments, can improve event coreference performance. Argument specific information propagation further help finding new relations. In the future, we propose to implement propagation based on temporal and other types of event relations.

8. Acknowledgments

This research was supported in part by DARPA grant FA8750-12-2-0342 funded under the DEFT program. Any opinions, findings and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of DARPA or the US government. We gratefully acknowledge anonymous reviewers for their valuable comments.

References

Ahn, D. (2006). The stages of event extraction. *ARTE '06 Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, 1–8.

Bagga, A., & Baldwin, B. (1998). Algorithms for scoring coreference chains. *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, 563–566.

Bagga, A., & Baldwin, B. (1999). Cross - Document Event Coreference: Annotations, Experiments, and Observations. *ACL-99 Workshop on Coreference and Its Applications*.

Bejan, C., & Harabagiu, S. (2010). Unsupervised event coreference resolution with rich linguistic features. *Proceedings of the 48th Annual Meeting of the Association for the Association for Computational Linguistics*(July), 1412–1422.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Chen, Z., & Ji, H. (2009). Graph-based event coreference resolution. *Proceedings of the 2009 Workshop on*

Type (counts)	Feature Name	Description
Discourse (5)	Sentence Distance	The number of sentences between the two events mentions.
	Event Distance	The number of event mentions between the two event mentions.
	Position	Whether one event is in the title, or first sentence.
Lexical (12)	Event String Similarity	Various string similarity measures are used to capture similarities between the event mention headwords, including Dice coefficient, edit distance, Jaro coefficient, lemma match and exact phrase match.
	Modifier Similarity	Similarities of the modifier words of two event mentions using Dice coefficient.
Syntactic (38)	Part Of Speech	Binary features for plurality, tense, noun or verb for the head words pairs of two event mentions.
	Syntactic Dependency	We add a feature if two event mentions connected by a syntactic type (one for each type).
	Modifier Features	Whether the two event mentions are modified; Dice coefficient of the pair of modifiers (if both exist); Whether two mention head words are both modified by negation.
	Determiner	Existence of determiner of the event mentions.
Semantic (16)	Event as Entity	Some nominal events are resolved by entity coreference engine, this is added as a boolean feature. Source of coreference comes from the Stanford Entity Coreference Engine (Lee et al., 2011) and the propagation steps.
	WordNet Similarity	Wu-Palmer WordNet similarity (Pedersen et al., 2004) is used for the event mention pairs.
	Senna embeddings	The Senna (Collobert et al., 2011) system created a word embeddings that maps words to a lower dimension vector space. Cosine similarity of event mention head words using this embedding is used as a feature.
	Distributional semantic	We use a semantic database to extract distributional semantic similarity between two events, which is described in Goyal et al. (2013).
	Verb Ocean	Whether the head of two event mentions have a specific relation in the Verb Ocean corpus (Chklovski & Pantel, 2004).
	Semantic Frame	This binary feature will be true if two event mentions trigger the same semantic frame (extracted using Semafor (Das et al., 2010)).
	Mention Type	IBM Sire's annotation (Florian et al., 2010) contains fine-grained mention type (such as attack events). A binary feature is added if two event mentions have the same Sire type.
Semantic (arguments) (34)	Argument similarities	Binary features indicating argument existence. There are also a number of features indicating similarities between arguments, including entity coreference of arguments (Source of coreference comes from the Stanford Entity Coreference engine (Lee et al., 2011) and the propagation steps.); Argument surface similarities using Dice coefficient; WuPalmer WordNet similarity between argument head words; Whether the number associated with each argument (e.g. 12 in *12 Somali*) are identical.
	Location	Surface match features include Dice and exact string match between mentions. Others include Location containment, alternatives name match, and mentions coreference. Location containment and alternative names are extracted using resources such as DBpedia (Lehmann et al., 2014) and Geonames.

Table 4: List of features (with counts) in the pairwise model.

Graph-based Methods for Natural Language Processing(August), 54–57.

Chen, Z., Ji, H., & Haralick, R. (2009). A pairwise event coreference model, feature impact and evaluation for event coreference resolution. *Proceedings of the Workshop on Events in Emerging Text Types*(3), 17–22.

Chinchor, N., & Sundheim, B. (1993). Muc-5 evaluation metrics. *Proceedings of the 5th conference on Message understanding*, 69–78.

Chklovski, T., & Pantel, P. (2004). VERB OCEAN : Mining the Web for Fine-Grained Semantic Verb Relations. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, 33–40.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011, November). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, *999888*, 2493–2537.

Cybulska, A., & Vossen, P. (2012). Using Semantic Relations to Solve Event Coreference in Text. *SemRel2012 in conjunction with LREC2012*, 60–67.

Danlos, L. (2003). Event coreference between two sentences. *Computing Meaning*, *77*, 271–288.

Das, D., Schneider, N., Chen, D., & Smith, N. (2010). Probabilistic frame-semantic parsing. *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference (NAACL 2010)*, *3*.

Elkhlifi, A., & Faiz, R. (2009). Automatic annotation approach of events in news articles. *International Journal of Computing & Information Sciences*, *7*(1), 40–50.

Finkel, J., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Florian, R., Pitrelli, J. F., Roukos, S., & Zitouni, I. (2010). Improving mention detection robustness to noisy input. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 335–345.

Goyal, K., Jauhar, S. K., Sachan, M., Srivastava, S., Li, H., & Hovy, E. (2013). A Structured Distributional Semantic Model for Event Co-reference. *Proceedings of The 51st Annual Meeting of the Association for Computational Linguistics*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, *11*(1), 10–18.

Hovy, E., Mitamura, T., Verdejo, F., Araki, J., & Philpot, A. (2013). Events are Not Simple: Identity, Non-Identity, and Quasi-Identity. *The 1st Workshop on EVENTS: Definitin, Detection, Coreference and Representation, NAACL-HLT 2013 Workshop*, 21–28.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 28–34.

Lee, H., Recasens, M., Chang, A., Surdeanu, M., & Jurafsky, D. (2012). Joint Entity and Event Coreference Resolution across Documents. *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., … Bizer, C. (2014). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*.

Li, P., Zhu, Q., & Zhu, X. (2011). A Clustering and Ranking Based Approach for Multi-document Event Fusion. *12th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, 159–165.

Luo, X. (2005). On coreference resolution performance metrics. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 25–32.

McConky, K., Nagi, R., Sudit, M., & Hughes, W. (2012, March). Improving event co-reference by context extraction and dynamic feature weighting. *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, 38–43.

Mendes, P., Jakob, M., Garcia-silva, A., & Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. *Proceedings of the 7th Conference on Semantic Systems*, 1–8.

Naughton, M. (2009). *Sentence Level Event Detection and Coreference Resolution*. Unpublished doctoral dissertation, National University of Ireland, Dublin.

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*(July), 104–111.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). Wordnet::similarity: measuring the relatedness of concepts. *Demonstration Papers at HLT-NAACL 2004*, 38–41.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*.

Pradhan, S., Ramshaw, L., Weischedel, R., MacBride, J., & Micciulla, L. (2007, September). Unrestricted Coreference: Identifying Entities and Events in OntoNotes. *Proceedings of the International Conference on Semantic Computing (ICSC 2007)*, 446–453.

Recasens, M., & Hovy, E. (2011). Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, *17*, 485–510.

Recasens, M., Hovy, E., & Martí, M. A. (2011). Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, *121*(6), 1138–1152.

Sangeetha, S., & Arock, M. (2012). Event Coreference Resolution using Mincut based Graph Clustering. *International Journal of Computing & Information Sciences*, 253–260.

Tratz, S., & Hovy, E. (2011). A fast, accurate, non-projective, semantically-enriched parser. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*(2010), 1257–1268.