Semantic Relations between Events and their Time, Locations and Participants for Event Coreference Resolution

Agata Cybulska
VU University Amsterdam
De Boelelaan 1105
1081HV Amsterdam

`a.k.cybulska@vu.nl`

Piek Vossen
VU University Amsterdam
De Boelelaan 1105
1081HV Amsterdam

`piek.vossen@vu.nl`

Abstract

In this study, we measure the contribution of different event components and particular semantic relations to the task of event coreference resolution. First we calculate what event times, locations and participants add to event coreference resolution. Secondly, we analyze the contribution by hyponymy and granularity within the participant component. Coreference of events is then calculated from the coreference match scores of each event component. Coreferent action candidates are accordingly filtered based on compatibility of their time, locations, or participants. We report the success rates of our experiments on a corpus annotated with coreferent events.

1 Introduction

In this paper, we present an approach to event coreference resolution that employs the importance of full and partial linguistic coreference between events and their participants, times and locations. The goal of this work is to measure the contribution of different components of event descriptions to the task of event coreference resolution. Another goal is to calculate what semantic relations add to event coreference. Considering the goals, we deliberately do not use machine learning as we want to have a clear picture of what the contributions are by different factors. Having an idea of how various event components influence event coreference, will guide the feature choice for machine learning.

Descriptions of one and the same event can differ in specificity and granularity (compare: *two students taken hostage in Beslanian school* vs. *two people taken hostage in a classroom in Beslan Russia*). High level events, as *war*, are more general and abstract with longer time span and group participants; low level events, e.g. a *shooting* event, are rather specific with shorter duration, and individual participants (Cybulska, Vossen, 2010). To capture differences between event representations and to identify relations between events, we applied an event model that consists of 4 components: a location, time, participant and an action slot (see Van Hage et al., 2011 for the formal SEM model along the same lines). In our previous work we extracted conflict-related actions (e.g. *war, genocide, shooting* or *fighting*) and their participants, locations and times from text. Next, we determine relations between event mentions, starting with getting some insights into event coreference.

2 Related Work

One of the recent approaches to event coreference resolution was proposed by Bejan and Harabagiu (2010), who experimented with nonparametric Bayesian models. Another one, by Chen et al. (2011) employs support vector machines with tree kernels and spectral graph partitioning. These approaches do not explicitly account for partial coreference of events, where some of the event components are related through hyponymy or part-of relationship, which is the focus of our work. Bejan and Harabagiu noted in their paper that not accounting for partial coreference is the reason for one of the common errors in their output. The approach of Chen et al. accounts for synonymy between mentions but not for meronymy or hyponymy.

Soft matching was successfully used for entity coreference resolution. Taxonomy based semantic similarity and semantic relatedness (Wikipedia based) were used as features in a machine learning approach to entity coreference by

156

Ponzetto and Strube (2006). Some semantic features based on synset relations in WordNet are used by Ng and Cardie (2002) and Ng (2005), while Harabagiu et al. (2001) use hyponymy, meronymy and other semantic relations from WordNet for NP coreference. They employ WordNet to distinguish between individuals and groups amongst entities of category PERSON.

Entity coreference has been used explicitly for event coreference resolution in the experiments by Lee et al. (2012); where entities and event clusters are merged by means of linear regression. Partial coreference is incorporated by using distributional similarity as one of features for cluster comparison. Other approaches use entities for event coreference in a more indirect way e.g. Bejan and Harabagiu (2008 and 2010) by using semantic roles as features for their SVM classifiers. Bejan and Harabagiu (2010) account only for synonymy amongst heads of semantic roles. Chen and Ji (2009) check for verbal argument compatibility for *Time-Within* and *Place* roles. Their results indicate that features related to event arguments only slightly (ca. +1% MUC and B3) improve event coreference, possibly due to wrong argument labeling. In this work, we measure the influence of time, place and participants on the task of event coreference resolution.

A theory-oriented discussion about the nature of full-, near- and non-identity and a continuum approach to entity coreference is presented in Recasens et al. (2011a). A discussion of full and quasi identity of events, pointing out the significance of partial coreference for coreference resolution, is held in Hovy et al. (2013).

Semantic shifts have been used before in NLP applications. Mulkar-Mehta et al. (2011a) investigated granularity shifts and structures in natural language. They focused on modeling part-whole relations between entities and events and causal relations between coarse and fine granularities. In their follow-up work (2011b), they described an algorithm for extracting causal granularity structures from text and its possible applications. Howard and Abramson (2012) use granularity types for prediction of rhetorical relations. Their results show that granularity types significantly improve prediction of rhetorical relations amongst clauses. In our work, we measure the contribution of shifts in granularity and abstraction to the task of event coreference resolution.

3 Approach to Coreference Resolution

Our approach to event coreference makes two crucial assumptions. First of all, we assume that solving coreference between actions is not enough to solve event coreference. If one only considers the action component it is impossible to determine whether two action mentions refer to the same event in reality, compare: *car bombing in Madrid in 1995* with *car bombing in Spain in 2009*. This is why, to solve event coreference we employ an event model which consists of 4 components: action, (human) participant(-s), location, and time. In accordance with the Quinean theory (1985), we assume that coreference between elements of the contextual setting of events is crucial for solving event coreference. Time and place in which an event happened form the starting point for event coreference resolution, compare: *genocide in Srebrenica* with *genocide in Rwanda*. Without time and place information event actions are just denotations of abstract classes of concepts. They need to be anchored in time and space to become instantiated.[1] Coreference thus only makes sense for events within the same time and place. Hence for each event mention in text, one should first define time and place and after that, for events occurring within a compatible time and space, search for linguistic coreference clues. From a practical point of view, determining event time and place should limit the number of candidates for coreferent events and improve the precision of event coreference resolution.

Secondly, we make the assumption that (linguistic) coreference is not an absolute notion. For example, *shooting* and *several shots* can refer to the same event and people may have different or vague intuitions about their identity (for a discussion of full and partial coreference see also Hovy et al. 2013). This approach employs a gradable notion of confidence in coreference with a continuum of non-disjoint events on which coreference of events (*bombing* vs. *bombing attack*) gradually transitions into other event relations as scriptal (event vs. its subevent e.g. *explosion* as a step in the script of a *bombing attack*), is-a (*bombing* being a kind of *attack*) and membership relations (*attack* being a member of *series of attacks*). The gradual notion of confidence in coreference inversely correlates with semantic distance between two instances. Semantic distance between instances of an event component can be determined by the kind of se-

---

[1] An interesting exception are event descriptions that depict instances of events that over time have become proper names as *World War II, 9/11, Srebrenica massacre*.

Event Components	Is-a: *Class>Subclass*	Inclusion: *Part-of, Member*
Location	*city>capital*	*Bosnia>Srebrenica*
Participants	*officer>colonel*	*army>soldier*
Time	*weekday>Friday*	*week>Monday*
Action	*attack>bombing*	*series of attacks>attack*

Table 1. Examples of event components related through hyponymy and meronymy.

mantic relation between them. In text one comes across specific and general actions, participants, time expressions and locations; compare e.g. *shooting, fighting, genocide* and *war,* or participants: *soldier* vs. (multiple) *soldiers* vs. *troops* and *multiple troops*. The same holds for time markers as *day, week* and *year* and for locations: *city* vs. *region* vs. *continent*. Table 1 exemplifies instances of event components related through hyponymy and meronymy. Mentions of event components are either (partially) overlapping or disjoint. Next to rather clear indicators typically used in coreference resolution as repetition, synonymy, anaphora and disjunction (negative indicator), significant relations between event components are along a hyponymy axis: class vs. its subclass such as *officer* being a subclass of the class *person*, instance-of a class such as *Bosnia* being an instance of the class *country*; and along a meronymy axis: member vs. group i.e. *Colonel Karremans* being a member of the group of *Dutch UN soldiers* or part vs. whole relation as *Srebrenica* being a part of *Bosnia*. For a thorough description of the model that captures the relationship between different semantic relations and coreference on one end of the spectrum and (if not disjoint) other event relations on the other, see our previous work (Cybulska, Vossen, 2012).

Within this approach, we analyze semantic relations and semantic distance between two instances of each event component, to obtain a coreference score per component. We do not only take exact lemma-based matches of event mentions into account but we allow for soft matching based on shifts in levels of granularity and abstraction. Our intuition is that shifts vs. agreement in the level of granularity and in the level of abstraction play a crucial role in establishing coreference relations; obviously together with other coreference indicators such as lemma repetition, anaphora, synonymy and disjunction. Once semantic distance and granularity agreement is calculated for every component of an event pair, the separate scores are combined into a single score for an event pair indicating the likelihood of real world coreference as a whole. Through empirical testing, we determine thresholds for establishing optimal coreference rela-

tions across events and their components.

4   Experiments

For the experiments we used the stand-off annotation of events (Lee et al. 2012) on top of the EventCorefBank (ECB) corpus[2], annotated with cross - document coreference between event mentions. The corpus contains 482 texts from Google News (selected based on inclusion of keywords such as *commercial transaction, attack, death* or *sports*) and grouped into 43 topics.

To measure the influence of time, location and participants on event coreference resolution, we first extract the set of events from the evaluation data. The ECB texts were processed by means of tools developed within the KYOTO project[3]. First, the corpus was lemmatized and tagged with PoS and syntactic information (Stanford Parser[4]). Next, word sense disambiguation was performed and the corpus was annotated with synsets from the English Wordnet (version 3.0) and with predefined ontology classes. The event ontology was manually assigned to 266 hypernyms in WordNet. It consists of four main semantic classes of concepts – one for each event component – location, time, participant and action which altogether cover 53964 synsets. All manually annotated actions from the corpus were used as input in the experiments. To extract participants, locations and times newly created extraction rules for English were used, based on manual annotation of event components in 5 independent texts. By means of the Kybot module of KYOTO, event times, participants and locations were extracted through rules employing some syntactic clues, PoS and combinatory information together with semantic class definition and exclusion by means of WordNet (Cybulska, Vossen, 2011).

There are two main stages to this experiment. First we generate preliminary chains of

---

[2] http://faculty.washington.edu/bejan/data/ECB1.0.tar.gz, Bejan and Harabagiu, 2010
[3] The ECB corpus texts after processing with the KYOTO tools (a pipeline of linguistic processors) are available at http://www.newsreader-project.eu/results/data/.
[4] http://www-nlp.stanford.edu/software/lex-parser.shtml

coreferring actions within a topic based on semantic similarity with the objective to ensure maximal recall. Similarity between mentions can be calculated by means of different techniques. We employed a taxonomy based edge counting technique of Leacock and Chodorow (1998) [5], which considers the closest hyponymy path in WordNet between two synsets scaled by the overall depth of the taxonomy:

$$(Si,j)=log(M(Di,j)/(2*Avg(Ddepth)))$$

where Si,j is the similarity between mentions i and j from M (total set of mentions in a topic); where $M(Di,j)$ is the minimal distance between two concepts and $Avg(Ddepth)$ is the average depth in WordNet for all meanings of all candidates in the topic. Mentions with relatively short semantic distance between their heads, constitute candidates for coreference chains. For mentions that use the same word, we ignore the synset but consider distance of 1. For synonyms, we use distance of 2. In all other cases, we add the hypernym distance to the initial value of 2. After obtaining the similarity scores for all mentions in a topic we normalize the scores. We created a matrix between all mentions in a topic and calculated the Leacock and Chodorow similarity (from now on also referred to as L&C) scores. A maximum recall was obtained if we keep equivalence relations for similarity scores of 20% or more of the highest score within a topic (usually the lemma). For each event mention, we thus keep candidate coreference relations to other mentions if the score is 0.2 or higher.

In our previous work, coreference of event actions was based solely on action similarity. In this part of the research, a second step was added to the process namely additional filtering of semantically similar actions based on compatibility of their participants, times and locations.

To experiment with semantic relations we use two different heuristics to determine participant compatibility: hyponymy and granularity. Note that this participant compatibility is not limited to full identity of participants. Soft matching of participants is more appropriate for the purpose of this task to account for cases of metonymy, e.g. *US aircrafts* instead of *US army.*

To generate chains of coreferent participants based on hyponymy, again we use the L&C (the same procedure as in case of action similarity). We determined the optimal coreference threshold for participant mentions on 0.7 normalized L&C score.

Our second heuristic calculates distance in granularity. Coreference chains are created in case of small distance in granularity levels between mentions. To determine granularity levels, we defined two semantic classes over synsets in WordNet: *gran_person* (e.g. *soldier, doctor*) denoting individual participants and *gran_group* referring to multiple participants (e.g. *army* or *hospital*). These two classes cover 36 WordNet hypernyms which map to 9922 synsets. On top of agreement in granularity levels, we also account for lexical granularity clues within a level such as number and multiplications. At this point we make a rough distinction between one and multiple items within a concept type (e.g. *gran_person*). Difference in granularity level or number is treated as indication of a granularity shift and is turned into a distance measure. To better handle 43415 [6] participant mentions that were POS - tagged as named entities, we decided to add an intermediate *gran_instance* class (for named entity participants that have no synsets such as person or organization names as *John,* or *Doctors Without Borders*) so that we can encourage number matching for our measurements of what granularity exclusively can contribute to event coreference. For agreement in semantic class level, two participant instances can maximally get 3 points. If there is 1 level difference between them (*gran_person > gran_instance* or *gran_instance > gran_group*) distance of 2 is determined. In case of participant pairs with *gran_person* and *gran_group* we have distance of 1. For number agreement we can maximally assign 2 points. If there is number disagreement – we assign 1 point. If there is both – level type agreement as well as number agreement a participant pair is given the maximum of 5 points.

As this paper aims at measuring the influence of different event components on event coreference, in the evaluation we filter our action chains based on location and time compatibility. In line with our theoretical approach, we see filtering on disjoint time and locations as crucial for event coreference resolution. For locations and time expressions, very strict thresholds were used, to avoid matches as *Monday* and *Tuesday*, sharing a short path in the taxonomy and consequently a high L&C score. The same holds for the granularity and domain heuristics. This is why, for the time being, only lemma and synonym matches are used. In the future we will look into treating proper names differently, and apply

---

[5] In the future we will also experiment with other methods.

[6] Out of the total of 54236 extracted participant mentions.

Heuristic	Event Slot	MUC			B3			CE AF	BLANC			Co NL L
		R	P	F	R	P	F	R/ P/F	R	P	F	F
LmB	All N&V	63. 8	82. 8	71. 2	65. 3	90. 6	75. 0	65. 9	68. 0	84. 1	71. 1	70. 7
L&C	Action	69. 4	72. 4	69. 5	69. 4	73. 3	68. 9	58. 7	68. 6	71. 8	67. 5	65. 2
Action L&C, Time Lm	Action Time	66. 0	77. 7	70. 6	66. 9	84. 2	73. 6	63. 9	68. 4	78. 1	70. 1	69. 4
Action L&C, Location Lm	Action Location	66. 3	77. 4	70. 6	67. 4	83. 0	73. 4	64. 1	68. 6	77. 3	70. 0	69. 3
Action L&C, Participant Lm	Action Participant	66. 0	78. 4	70. 8	67. 0	84. 9	73. 9	64. 5	68. 6	79. 0	70. 4	69. 7
Action L&C, ParticipantL&C	Action Participant	65. 2	79. 4	70. 7	66. 8	85. 7	74. 1	64. 9	68. 5	79. 7	70. 4	69. 8
Action L&C, Part.granularity	Action Participant	66. 5	0,7 7.8	70. 4	67. 6	81. 7	72. 2	62. 5	68. 3	77. 9	69. 4	68. 2

Table 2. Coreference Evaluation in MUC, B3, CEAFm, BLANC and CoNLL F (macro averages).

similarity and granularity measurements to time expressions and locations that are not proper names. We will also consider employing geo and temporal ontologies containing proper names.

Our current approach boosts the score of action coreference for each participant, time and location coreference chain they share, taking the coreference score of each chain as a weight for sharing. We used a formula in which membership to a coreference set of an event is initially based on the coreference score of the action mention but it is strengthened by the proportion that participants, time references or locations are shared with other mentions:

$Coref(m,E)=MAXLC(m,E) + P(p) \lor P(t) \lor P(l)$

where E is the set of mentions in action coreference set, $MAXLC$ is the highest similarity score for the mention m in the set E. The coreference score of action mention m equals the sum of the maximum coreference score $MAXLC$, and proportion P of overlapping participants p (of m with the other members of the set) or times t or locations l, with other members of the set.

5   Evaluation Results

For the evaluation, the manual annotations of actions from the ECB corpus were used as key chains and were compared with the response chains generated for each topic by means of the above described heuristics. Since our goal was to evaluate the importance of coreference between other event components (than actions) for the task of event coreference resolution, we compare our evaluation results with system results based on action similarity only, i.e. when disregarding other event components. We also aimed at getting some insights into the contribution by shifts in hyponymy and granularity (soft matching). This is why we use a lemma baseline *(LmB)* that assigns coreference relation to all nouns and verbs that belong to the same lemma (strict matching). Table 2 presents coreference evaluation results achieved by means of the different heuristics: the L&C measure, granularity agreement as well as lemma match *(Lm)* in comparison to the baseline results *(LmB)* in terms of recall *(R),* precision *(P)* and F-score *(F),* employing the commonly used coreference evaluation metrics: MUC (Vilain, 1995), B3 (Bagga, Baldwin, 1998), mention-based CEAF (Luo, 2005), BLANC (Recasens, Hovy, 2011b), and CoNLL F1 (Pradhan et al., 2011).

Compared to the lemma baseline, our approach using similarity of event actions only (second row in table 2), across majority of the evaluation metrics improves R with up to 6% while loses (2-17%) $P,$ what is expected. It is worth noticing, that the baseline achieves remarkably good results, what could be caused by the fact that the annotators are drawn to pick up on the most obvious coreference cases. Within narrowly defined topics, such as news articles of the same day on a specific event, these are usually expressed by the same lemma.

When comparing the contribution of participants, times and locations (all lemma matches for the sake of comparison) with the approach using

exclusively action similarity, we see that the approach combining action and participant components achieved slightly better results (ca. 1% higher precision scores) than the two other approaches employing time and location slots. Altogether, the differences between the scores are in this case rather subtle. When analyzing these results one must keep in mind that these evaluation scores are conditioned by the fact that participant descriptions occur much more frequently in event descriptions than time and place markers. [7]

Out of the two different heuristics used in participant approaches; ca. 1% higher F-scores (a 2-4% improvement of precision) on most evaluation metrics were obtained with L&C similarity. Both participant approaches in most metrics improve the F-scores achieved by the action similarity heuristic; the granularity approach with ca. 1-4% and participant similarity with ca. 1-6%.

Compared to the lemma baseline *(LmB)*, our best scoring approach of all, that is action similarity with participant similarity, on most metrics loses ca. 1% on the F-scores. It gains up to 2 points in recall, while generating output with ca. 4% lower precision. This small decline in F measure can be motivated by the fact that we are dealing here with within topic coreference (although cross – document). Also, evaluation data seem to be biased towards coreference chains around smaller events. Corpora, even those annotated with cross-document coreference of events, (intentionally) tend to be composed around specific real world events, such as attacks or earthquakes, so that coreference chains are captured in a rather small time frame. The diversity of event instances from the same type of event class that happened in different time frames, places and with different participants is much lower in such a corpus than in the real world, e.g. realistic daily news streams. The relatively high scores achieved by the lemma baseline show the need for different event coreference datasets, where cross-document coreference is marked in text across different instances of particular event classes, e.g. describing two different *wars* that take place over longer stretches of time and include similar types of events. Only then the data will become more representative of the sampled population.

Compared to evaluation results achieved in related work:

- Bejan and Harabagiu, 2010: 83.8% B3 F,

76.7% CEAF F on the ACE (2005) data set and on the ECB corpus 90% B3 F, 86.5% CEAF F-score
- Lee et al., 2012: 62.7% MUC, 67.7% B3 F, 33.9% (entity based) CEAF,71.7% BLANC F-score on the ECB corpus
- Chen et al., 2011: 46.91% B3 F on the OntoNotes 2.0 corpus

by means of our best scoring approach, using action and participant similarity, coreference between actions was solved with an F-score of 70.7% MUC, 74.1% B3, 64.9% CEAFm, 70.4% BLANC F and 69.8 CoNLL F1. Considered that our approach neither considers anaphora resolution nor syntactic features, there is definitely room for improvement of event coreference resolution for an approach that combines these with semantic matches of event components.

6 Conclusion and Future Work

In this paper, we presented our approach to event coreference that employs the importance of coreference (also partial linguistic coreference) between participants, locations and times for the task of event coreference resolution. Our results show that filtering coreferent action candidates based on compatibility of their participants (our best scoring approach) in comparison to the baseline slightly improves precision of the resolution of coreference between events. The results are especially promising given the limitations of the approach, such as not performing anaphora resolution. In the future, we will further experiment with coreference resolution, amongst others by applying our method to cross – topic coreference of events, to find out whether there is more variation in structural properties if one considers not only different texts, but also various topics. If that is the case, semantic matches should turn out to be even more important.

Furthermore, we will experiment with clustering techniques as a heuristic to identify coreference sets, where different event components as well as hyponymy and meronymy agreement, are used as features.

---

[7] From the ECB corpus we extracted 54236 participant, 5728 location and 3435 time mentions.

References

ACE-Event. 2005. ACE English Annotation Guidelines for Events, ver. 5.4.3 2005.07.01.

Bagga, Amit and Breck Baldwin, "Algorithms for Scoring Coreference Chains", in Proceedings of LREC 1998

Bejan, Cosmin Adrian and Sanda Harabagiu, "A Linguistic Resource for Discovering Event Structures and Resolving Event Coreference", in Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, 2008

Bejan, Cosmin Adrian and Sanda Harabagiu, "Unsupervised Event Coreference Resolution with Rich Linguistic Features", in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 2010

Chalupsky, Hans et al., "The RACR Machine Reading System", in prep. 2013

Chen, Zheng and Heng Ji, "Event Coreference Resolution: Algorithm, Feature Impact and Evaluation", in Proceedings of Events in Emerging Text Types (eETTs) Workshop, in conjunction with RANLP, Bulgaria, 2009

Chen, Bin, Su, Jian, Pan, Sinno Jialin and Chew Lim Tan, "A Unified Event Coreference Resolution by Integrating Multiple Resolvers", in Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, November, 2011

Cybulska, Agata and Piek Vossen, "Event models for Historical Perspectives: Determining Relations between High and Low Level Events in Text, Based on the Classification of Time, Location and Participants", in Proceedings of LREC 2010, Valletta, Malta, May 17-23, 2010

Cybulska, Agata and Piek Vossen, "Historical Event Extraction from Text", in Proceedings of ACL LaTeCH, Portland, US, June 2011

Cybulska, Agata, and Piek Vossen, "Using Semantic Relations to Solve Event Coreference in Text", in Proceedings of the Workshop: Semantic Relations-II. Enhancing Resources and Applications (SemRel2012), Istanbul, Turkey, May 2012

van Hage, Willem Robert, Malaisé, Véronique, Segers, Roxane, Hollink, Laura and Guus Schreiber, "Design and use of the Simple Event Model (SEM)", in Journal of Web Semantics 9(2):128-136, July 2011.

Harabagiu, Sanda M., Bunescu, Razvan C. and Steven J. Maiorano, "Text and Knowledge Mining for Coreference Resolution", in Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, 2001

Howald, Blake Stephen and Martha Abramson, "The Use of Granularity in Rhetorical prediction", in Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM), pages 44-48, Montreal, Canada, 2012

Hovy, Eduard, Mitamura, Teruko, Verdejo, Felisa and Philpot, Andrew, "Identity and Quasi-Identity Relations for Event Coreference", in prep.2013

Ide, Nancy and David Woolner, 2007, "Historical Ontologies", in: Ahmad, Khurshid, Brewster, Christopher, and Mark Stevenson (eds.), Words and Intelligence II: Essays in Honor of Yorick Wilks, Springer, 137-152.

Kuebler, Sandra and Denislava Zhekova, "Singletons and Coreference Resolution Evaluation", in Proceedings of Recent Advances in NLP, Hissar, Bulgaria, September, 2011

Leacock, Claudia and Martin Chodorow, "Combining local context with WordNet similarity for word sense identification", in Christiane Fellbaum (ed.), WordNet: A lexical Reference System and its Application, MIT Press, Cambridge, MA.

Lee, Heeyoung, Recasens, Marta, Chang, Angel, Surdeanu, Mihai and Dan Jurafsky, "Joint Entity and Event Coreference Resolution across Documents", Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL), 2012

Luo, Xiaoqiang, "On coreference resolution performance metrics", in *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP-2005)*, 2005

Mulkar-Mehta, Rutu, Hobbs, Jerry R. and Eduard Hovy,"Granularity in Natural Language Discourse",in Proceedings of International Conference on Computational Semantics, 2011a

Mulkar-Mehta, Rutu, Hobbs, Jerry R. and Eduard Hovy, "Applications and Discovery of Granularity Structures in Natural Language Discourse", in Proceedings of The Tenth International Symposium on Logical Formalizations of Commonsense Reasoning at the AAAI Spring Symposium, Palo Alto, 2011b

Ng, Vincent, "Machine Learning for Coreference Resolution: From Local Classification to Global Ranking", in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005

Ng, Vincent and Claire Cardie, "Improving Machine Learning Approaches to Coreference Resolution", in *Proceedings of the 40th Annual Meeting of the*

Association for Computational Linguistics (ACL), Philadelphia, 2002

Ponzetto, Simone Paolo and Michael Strube, "Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution", in Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 192-199, New York, 2006

Pradhan, Sameer, Ramshaw, Lance, Marcus, Mitchell, Palmer, Martha, Weischedel, Ralph and Nianwen Xue, "CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes", in *Proceedings of CoNLL 2011: Shared Task,* 2011.

Quine, Willard V., "Events and Reification" in E. LePore and B.P. McLaughlin (eds.), Action and Events, Basil Blackwell, New York, 1985

Recasens, Marta, Hovy, Eduard and M. Antònia Martí, "Identity, non-identity, and near-identity: Addressing the complexity of coreference", Lingua, 121(6):1138-1152, 2011a

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

Vilain, Marc, Burger, John, Aberdeen, John, Connolly Dennis and Lynette Hirschman, "A model theoretic coreference scoring scheme", in *Proceedings of MUC-6*, 1995