Translating Granularity of Event Slots into Features for Event Coreference Resolution.

Agata Cybulska
VU University Amsterdam
De Boelelaan 1105
Amsterdam, 1081HV
a.k.cybulska@vu.nl

Piek Vossen
VU University Amsterdam
De Boelelaan 1105
Amsterdam, 1081HV
piek.vossen@vu.nl

Abstract

Using clues from event semantics to solve coreference, we present an "event template" approach to cross-document event coreference resolution on news articles. The approach uses a pairwise model, in which event information is compared along five semantically motivated slots of an event template. The templates, filled in on the sentence level for every event mention from the data set, are used for supervised classification. In this study, we determine granularity of events and we use the grain size as a clue for solving event coreference. We experiment with a newly-created granularity ontology employing granularity levels of locations, times and human participants as well as event durations as features in event coreference resolution. The granularity ontology is available for research. Results show that determining granularity along semantic event slots, even on the sentence level exclusively, improves precision and solves event coreference with scores comparable to those achieved in related work.

1 Introduction

Event coreference resolution is the task of determining whether two event mentions refer to the same event instance. This paper explores cross-document resolution of coreference between events in a news corpus. We use granularity as an indication of event coreference. Our approach renders the semantic structure of event descriptions into arrangement of features for machine learning.

We use the granularity of events as a clue for event coreference resolution. The intuition behind this approach is, that an event with a longer duration, that happens on a bigger area and with multiple participants involved (for instance *a war between Russia and Ukraine*) might be related to but will probably not fully corefer with a "lower level" event of shorter duration and with single participants involved (e.g. *A Russian soldier has shot dead a Ukrainian naval officer*).

We experiment with an "event template" approach to event coreference resolution. The way in which event information can be semantically categorized is used in an event template to shape comparison of information about two event descriptions. Coreference between mentions of two events is determined through compatibility of slots of a pair of event templates. For the experiments, we use the ECB+ dataset (Cybulska and Vossen, 2014b). The five slots in our event template correspond to different elements of event information as annotated in the ECB+. The considered event slots are: 1) event action that is the event trigger (following the ACE (LDC, 2005) terminology) and four kinds of event arguments: 2) time, 3) location, 4) human and 5) non-human participant slots (for more information see Cybulska and Vossen (2014a)). An event template can be filled at different levels of information such as the entire document, a paragraph or a sentence. The approach investigated in this study operates at the sentence level which means that event templates are filled only with information available in the sentence in which an event mention occurs (for a report on experiments with a two step approach first considering document and subsequently sentence templates, see Cybulska and Vossen (2015)). Figure 1 considers an excerpt from topic one, text seven of the ECB corpus (Bejan and Harabagiu, 2010). Table 1 shows the distribution of

Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015, pages 1–10,
Denver, Colorado, June 4, 2015.

Event Slot	Sentence Template 1	Sentence Template 2
Action	*entered*	*headed*
Time	*N/A*	*on Tuesday*
Location	*Promises*	*to a Malibu treatment facility*
Human Participant	*actress*	*actress*
Non-human Participant	*N/A*	*N/A*

Table 1: Sentence templates ECB topic1, text 7, sentences 1 and 2.

The "American Pie" actress has entered Promises for undisclosed reasons. The actress, 33, reportedly headed to a Malibu treatment facility on Tuesday.

Figure 1: Topic 1, text 7, ECB (Bejan and Harabagiu, 2010).

event information over the five event slots (as annotated in the ECB+) in the two example sentences. In the event template approach different kinds of event information are contrasted per slot of the template (Table 3).

We determine the grain-size within slots of the event template. The idea is to represent the grain size of the event action as well as of the entities involved with it by means of granularity features. To capture granularity we employ durations of event actions (Gusev et al., 2011) and granularity levels of event participants, time and locations. To determine granularity levels, a new granularity ontology consisting of 15 semantic classes is used. The 15 predefined semantic classes represent different granularity levels, which are defined over 434 hypernyms in WordNet, covering 11979 WordNet synsets. We make the granularity ontology available for research.

This work sheds light on the task of cross-document resolution of coreference between mentions of events in text. This study explores the actual task of resolution of coreference between two event descriptions, without letting topic classifiers first solve most of event ambiguity (following the insights of Cybulska and Vossen (2014b)). The two main contributions of this study are: (1) a new granularity ontology of event participants, times and locations and (2) a new "sentence template" approach to event coreference resolution that solves event coreference along five slots of an event template. To the best of our knowledge granularity of locations, times and human participants of events as well as durations of event actions has not been used before to solve event coreference.

We will first take a closer look at the notion of granularity and the new granularity ontology in section 2. We delineate our approach in section 3. Section 4 reports on the experiments with the new method the results of which are compared with related work in section 5. We conclude in section 6.

2 Granularity

The notion of granularity was described by (Keet, 2008) as the ability to represent and operate on different levels of detail in data, information, and knowledge. *Granularity deals with organizing data, information, and knowledge in greater or lesser detail that resides in a granular level or level of granularity and which is granulated according to certain criteria, which thereby give a* [granular] *perspective (...) on the subject domain.* (Keet, 2008). A lower granularity level captures a more detailed data representation than a more abstract higher level, which leaves out some details.

People view the world at different granularities. Humans are able to switch among different granularities of world conceptualizations (Hobbs, 1985). In a reasoning process a granularity level is distinguished, depending on what is relevant for a particular situation. Hobbs presented a framework for a theory of granularity.

Few other researchers looked at granularity in natural language. Considered the variation in the degree of specification of word meaning, Mani (1998) suggested development of a knowledge representation, which makes the notion of granularity explicit. Mani applied shifts in granularity to problems of polysemy and underspecification of nominaliza-

eng-30-08160276-n,gran_group,"citizenry_1,people_2"
eng-30-10638385-n,gran_person,"spokesperson_1,interpreter_3,representative_2,voice_8"
eng-30-15235126-n,gran_second,"second_1,sec_1"
eng-30-15234942-n,gran_min,"quarter_4"
eng-30-15117516-n,gran_hr,"hours_2"
eng-30-15163005-n,gran_day,"day_of_the_week_1"
eng-30-15136147-n,gran_week,"week_3,calendar_week_1"
eng-30-15209706-n,gran_month,"Gregorian_calendar_month_1"
eng-30-15239579-n,gran_season,"season_1"
eng-30-15203791-n,gran_year,"year_1"
eng-30-15231415-n,gran_thousands_years,"Bronze_Age_1"
eng-30-03449564-n,gran_street,"government_building_1"
eng-30-08537837-n,gran_city,"city_district_1"
eng-30-08898002-n,gran_country,"Upper_Egypt_1"
eng-30-08699426-n,gran_continent,"East_Africa_1"

Figure 2: Example entries from the granularity ontology file.

tions. Change in granularity was considered as a special case of abstraction in which elements, which are indistinguishable in a particular context, are collapsed. Mani focused on grain-size shifts amongst polysemous events.

Mulkar-Mehta et al. (2011b) describe event granularity as the concept of breaking down a higher-level event into smaller parts, fine-grained events such that each smaller granule plays a part in the higher level whole. Relation types that can exist between the objects at coarse and fine granularity are part-whole relationships amongst entities and events, and causal relationships. Based on annotation of granularity relations in text, the authors conclude that part-whole and causal relations are a good indication of shifts in granularity.

In this study we focus on the notion of granularity in event descriptions. We present a new granularity ontology, which is an attempt at capturing grain-size of events explicitly for the purpose of usage in NLP applications. We use a taxonomy based ontology to distinguish between coarse- and fine-grained granularities of different parts of event descriptions. We apply shifts in granularity to resolution of event coreference. The motivation behind this approach is an expected correlation between agreement or disagreement in grain-size levels and the notion of coreference. Agreement or small granularity differences are expected to indicate coreference. Bigger distance in granularity is expected to be a negative indicator of coreference or to indicate other event relations as scriptal or event membership. In the experiments described in this paper, we let a machine learning algorithm learn the relationships between different granularities and the notion of coreference. To capture differences in grain-size of events we employ both: (1) conceptual granularity clues being a manifestation of granularity in the form of inherent properties of word meanings, as well as (2) lexical grain-size indication expressed in number and multiplication. The intrinsic, conceptual granularity is captured by means of a number of granularity levels defined in the granularity ontology. Furthermore, we use durations of events as indication of grain size for event actions.

2.1 Granularity Ontology

We focus here on partonomic granularity relations (representing granularity through the part-of relation) between entities and events. To establish granularities of event participants, times and locations we created a new granularity ontology. Semantic classes relating to granularity levels were defined over synsets in WordNet. In the experiments we employ granularity levels to capture granularity agreement and shifts amongst event participants, times and locations. Our 15 semantic classes belong to four relationships from the taxonomy of meronymic

relations by Winston et al. (1987). Granularity levels of the human participant slot are contained within Winston's et al. Member-Collection relations. Our temporal granularity levels make part of Winston's Portion-Mass relationships and our locational levels are in line with Place-Area relations in Winston's taxonomy.

Figure 2 presents a fragment of the granularity ontology with synset examples for every ontology class. The file is comma separated. In the first column synsets from WordNet 3.0 are indicated. In the second column the granularity levels are captured and the third one indicates the synset IDs as stored in the Natural Language Toolkit (NLTK, (Bird et al., 2009)). The choice of the 15 granularity classes was motivated by an analysis of event descriptions in the news. We intended to capture shifts in granularity that seemed meaningful for event coreference resolution on a news corpus such as the ECB or ECB+. We manually assigned the semantic classes to 434 hypernyms in WordNet which are linked to 11979 synsets. We recognize a number of granularity levels per event slot: nine grain levels for time expressions, four for locations and two for human participants, as presented in Table 2.

2.2 Lexical Granularity Clues

On top of granularity levels, we also account for lexical granularity clues within a level such as number indication and multiplications. At this point we only make a distinction between *single* and *multiple* "items" within a concept type (based on POS clues and occurrence of multiplications). Three kinds of parts of speech are used to determine number of a mention: (1) nominal tags: *NN, NNS, NNP, NNPS*, (2) personal pronouns tagged by the NLTK's default POS tagger as *PRP* and (3) numbers with tag *CD*. For instance the phrase *twenty soldiers* is POS-tagged as follows: *[('20', 'CD'), ('soldiers', 'NNS')]*. The nominal POS tag *NNS* is considered to indicate plural nouns. Additionally, if there is a number indication in a mention (POS-tag *CD* and lemma other than *one*), the phrase would be assigned plural number by default. If there are multiple nouns in a mention, we assign the number of the majority of nouns. If there is a tie, the number of the last noun in a mention would be decisive. For example *[('20', 'CD'), ('soldiers', 'NNS')]* would be assigned the granularity level *gran_person* and number *plural*. While *one soldier* would trigger the following analysis: *[('one', 'CD'), ('soldier', 'NN')]*, also assigned the granularity level *gran_person* but number *singular*. Since there are often multiple instances of an event slot in the sentence, there can be multiple granularity levels to consider. We calculate cosine similarity of granularity and number indications per event slot (if instantiated in the sentence) for two compared events. In the future, we will experiment with expressing the grain-size by means of numeric estimates of number of participants, duration and size of an area on which an event happened, e.g. indicating that the Boston area is ca. 125 km2 and the country of France of ca. 551500 km2.

2.3 Event Durations

To capture granularity of event actions (in Winston et al. (1987) Feature-Activity relation) we employ duration distributions from the database of event durations by Gusev et al. (2011). The lexicon of event durations (`http://cs.stanford.edu/people/agusev/durations/`) captures durations for events (with or without syntactic objects) inferred by means of web query patterns. Duration distributions were learned with an unsupervised approach. Eight duration levels are considered: *seconds, minutes, hours, days, weeks, months, years* and *decades*. The durations database covers the 1000 most frequent verbs with 10 most frequent grammatical objects of each verb from a newspaper corpus from the New York Times. For our granularity experiments we used duration distributions as determined for these 10000 events. A binary feature indicates whether there is overlap in most frequent duration levels of two events. Currently, since our approach does not consider syntactic dependencies, the duration feature is specified when disregarding the syntactic objects.

3 The Approach

We experimented with a decision-tree (hereafter also *DT*) supervised pairwise binary classifier to determine coreference of pairs of event mentions represented through templates filled in at the sentence level. We run preliminary experiments with a linear SVM and a multinomial Naive Bayes classifier

Event Slot	Granularity Class	Description	Synset Example
Human Participant	*gran_person*	individuals	spokesperson_1
	gran_group	groups or organizations	people_2
Location	*gran_street*	areas up to the size of a building	government_building_1
	gran_city	city districts and cities	city_district_1
	gran_country	size of a country	Upper_Egypt_1
	gran_continent	size of multiple countries	East_Africa_1
Time	*gran_second*	duration up to a minute	sec_1
	gran_min	from a minute to an hour	quarter_4
	gran_hr	from an hour up to 24 hours	hours_2
	gran_day	one to few days, less than a week	day_of_the_week_1
	gran_week	one to few weeks, less than a month	calendar_week_1
	gran_month	indication on the month level	Gregorian_calendar_month_1
	gran_season	few months	season_1
	gran_year	one or multiple years	year_1
	gran_thousands_years	thousands of years	Bronze_Age_1

Table 2: Granularity ontology classes.

Template Slot		Feature	Explanation
Action	Active mention coreference is solved for	Lemma overlap (L)	Numeric feature: overlap percentage.
		Duration overlap (G)	Binary: overlap in most frequent level.
		Synset overlap (S)	Numeric: overlap percentage.
		Discourse location (D)	Location within discourse. Binary:
		- document	- the same document or not
		- sentence	- the same sentence or not.
	Other sentence mentions	Lemma overlap (L)	Numeric: overlap percentage.
		Synset overlap (S)	Numeric: overlap percentage.
Location		Lemma overlap (L)	Numeric: overlap percentage.
		Granularity & num. overlap (G)	Numeric: cosine similarity.
		Synset overlap (S)	Numeric: overlap percentage.
Time		Lemma overlap (L)	Numeric: overlap percentage.
		Granularity & num. overlap (G)	Numeric: cosine similarity.
		Synset overlap (S)	Numeric: overlap percentage.
Human Participant		Lemma overlap (L)	Numeric: overlap percentage.
		Granularity & num. overlap (G)	Numeric: cosine similarity.
		Synset overlap (S)	Numeric: overlap percentage.
Non-Human Participant		Lemma overlap (L)	Numeric: overlap percentage.
		Synset overlap (S)	Numeric: overlap percentage.

Table 3: Features used in the experiments grouped into four categories: L - lemma based, G - granularity and number, D - discourse and S - synset based features.

but the decision-tree classifier outperformed both of them. We trained the DT classifier on an unbalanced training set of positive and negative samples.

In the experiments different features were assigned values per event slot. Table 3 presents all features that we experimented with. The lemma overlap feature (L) expresses a percentage of overlapping lemmas between two instances of an event slot (after removal of skip words), if instantiated in the sentence. Features indicating granularity and number compatibility of an event slot (G), are specified for every location, time and human participant mention in the sentence. Frequently, one ends up with multiple entity mentions from the same sentence for an action mention (the relation between an event and entities involved with it is not annotated in ECB+). To express the degree of overlap in grain size of mentions we used cosine similarity. For the action slot overlap in duration level of the active mentions is considered as a binary feature. For all five slots a percentage of synset overlap is calculated (S). Finally there are two features indicating mentions location within the discourse (D), specifying if mentions come from the same sentence or document.

Prior to being fed to the classifier, numeric feature vectors were normalized (missing values were imputed). We used grid search with ten fold cross-validation to optimize the depth of the decision-tree algorithm (entropy was used as the criterion).

Pairs of event templates were classified by means of the DT classifier when employing features from Table 3. To identify the final equivalence classes of corefering event mentions, mentions were grouped based on corefering pair overlap.

4 Experiments

4.1 Corpus

For the experiments we used the true mentions from the ECB+ corpus (Cybulska and Vossen, 2014b) which is an extended and re-annotated version of the ECB corpus (Bejan and Harabagiu, 2010). The ECB+ corpus contains a new corpus component, consisting of 502 texts, describing different instances of event types that were already captured by the 43 topics of the ECB.

As recommended by the authors in the release notes, for experiments on event coreference we used a subset of ECB+ annotations (based on a list of 1840 selected sentences), that were additionally reviewed with focus on coreference relations. Table 4 presents information about the data set used for the experiments. We divided the corpus into a training set (topics 1-35) and test set (topics 36 - 45).

4.2 Experimental Set Up

The ECB+ texts are available in the XML format. The texts are tokenized, hence no sentence segmentation nor tokenization needed to be done. We POS-tagged and lemmatized the corpus sentences. For the experiments we used tools from the Natural Language Toolkit (Bird et al., 2009)[1]: the NLTK's default POS tagger, and WordNet lemmatizer[2] as well as WordNet synset assignment by the NLTK[3]. For machine learning experiments we used scikit-learn (Pedregosa et al., 2011).

4.3 Singleton Baseline

As a baseline we consider event coreference evaluation scores generated taking into account all event mentions as singletons. In the singleton baseline response there are no "coreference chains" of more than one element. First row of Table 5 presents the singleton baseline results (BL) in terms of recall (R), precision (P) and F-score (F) by employing the coreference resolution evaluation metrics: MUC (Vilain et al., 1995), B3 (Bagga and Baldwin, 1998), mention-based CEAF (Luo, 2005), BLANC (Recasens and Hovy, 2011), and CoNLL F1 (Prad-

[1] NLTK version 2.0.4

[2] www.nltk.org/_modules/nltk/stem/wordnet.html

[3] http://nltk.org/_modules/nltk/corpus/reader/wordnet.html

ECB+ Corpus	#
Topics	43
Texts	982
Action mentions	6833
Location mentions	1173
Time mentions	1093
Human participant mentions	4615
Non-human participant mentions	1408
Coreference chains	1958

Table 4: ECB+ statistics.

Heuristic	Features	MUC			B3			CEAF	BLANC			CoNLL
		R	P	F	R	P	F	F	R	P	F	F
BL	-	0	0	0	45	100	62	45	50	50	50	39
DT	L	**43**	**77**	**55**	**58**	**86**	**69**	**58**	**60**	**69**	**63**	**64**
DT	LG	36	**77**	**49**	55	**90**	**68**	**56**	56	**74**	**60**	**60**
DT	LGD	28	77	42	52	93	67	55	55	77	58	57
DT	LGDS	16	76	27	49	96	65	52	52	68	54	50

Table 5: Sentence template approach to event coreference resolution evaluated on the ECB+ corpus in MUC, B3, mention-based CEAF, BLANC and CoNLL F in comparison to the singleton baseline BL.

Approach	Data	Model	MUC			B3			CEAF	BLANC			CoNLL
			R	P	F	R	P	F	F	R	P	F	F
BL	ECB+	-	0	0	0	45	100	62	54	50	50	50	39
B&H	ECB	HDp	52	90	66	69	96	80	71	NA	NA	NA	NA
Lee	ECB	LR	63	63	63	63	74	68	34	68	79	72	55
STA - L	ECB+	DT	43	77	55	58	86	69	66	60	69	63	64
STA - LG	ECB+	DT	36	77	49	55	90	68	63	56	74	60	60

Table 6: Best scoring STA approaches using feature sets L and LG evaluated in MUC, B3, entity-based CEAF, BLANC and CoNLL F; in comparison with related studies and the BL baseline. Note that the STA uses gold and related approaches system mentions.

han et al., 2011). When discussing event coreference scores must be noted that some of the commonly used metrics depend on the evaluation data set. This results in scores going up or down with the number of singleton items in the data (Recasens and Hovy, 2011). Our singleton baseline gives us zero scores in MUC, which is due to the fact that the MUC measure promotes longer chains. B3 on the other hand seems to give additional points to responses with more singletons, hence the remarkably high scores achieved by the baseline BL in B3. CEAF and BLANC as well as the CoNLL measures (the latter being an average of MUC, B3 and entity CEAF) give more realistic results.

4.4 Results

We evaluate the system output produced by the decision-tree classifier after merging pairs of event mentions with common elements into equivalence classes. The response chains generated with: (1) lemma feature set L, (2) lemma and granularity LG, (3) lemma, granularity and discourse LGD, and (4) lemma, granularity, discourse and synset features LGDS are evaluated in Table 5 in terms of R, P and F-score by employing the MUC, B3, mention-based CEAF, BLANC and CoNLL F1 metrics.

The highest F scores reached the event clusters created by the decision-tree classifier employing feature set L (marked in bold in the table). We observe a 13% improvement over the baseline BL in mention-based CEAF F and in BLANC F and a 25% gain in CoNLL F.

Addition of granularity features (LG) increases the precision scores in B3 and BLANC by 4-5%. The recall scores decrease but the F scores in most measures (except for MUC) are between 56-68%. Employing discourse with lemma and granularity features (LGD) gives us some extra precision points but costs us even more recall. Synset features lower precision and recall.

Note that these results were generated when disregarding syntactic roles and POS information. No anaphora resolution was performed and we did not group the corpus texts into topics before solving coreference between event mentions at the sentence level, which would significantly simplify the task (Cybulska and Vossen, 2014b). In the future we will run experiments aiming at improving the recall for instance through addition of semantic similarity features (in combination with the currently used fea-

tures). We will also investigate the influence of syntactic features on the results.

5 Related Work

Granularity shifts and structures were recently investigated in the context of NLP applications by Mulkar-Mehta et al. (2011b). In their follow-up work (Mulkar-Mehta et al., 2011a) they describe an algorithm for extracting causal granularity structures from text and its possible applications in question answering and text summarization.
Howald and Abramson (2012) successfully used granularity types as features for prediction of rhetorical relations with a 37% performance increase.

As for event coreference resolution, Humpreys et al. (1997) performed coreference merging between event template structures. Our event template however is much more restricted (five slots only) and it is filled and compared at the level of sentence while Humpreys et al. consider discourse events and entities for event coreference resolution. No coreference evaluation scores are reported.

Considering the limitations of the event coreference resolution measures, for the sake of a meaningful comparison, it is important to consider similar data sets. The ECB and ECB+ are the only available resources annotated with both: within- and cross-document event coreference. We were unable to run our experiments on the ECB corpus, because no specific entity types are annotated in the ECB and our work depends on those for granularity estimates.[4] To the best of our knowledge, no baseline has been set yet for event coreference resolution on the ECB+ corpus. Accordingly we will look at results achieved in cross-document event coreference resolution on the ECB corpus which is a subset of ECB+, and so the closest to the data set used in our experiments. For the sake of convenience, in Table 6 we compare the best results by the sentence template approach (when using lemma features *STA - L* and a combination of lemma and granularity features *STA - LG*) with the results achieved in related studies. *B&H* stands for the approach of Bejan and Harabagiu (2010) using HDp - hierarchical Dirichlet process and *Lee* refers to the approach of Lee et al. (2012) using LR - linear regression. *BL* denotes the results by the singleton baseline.

In comparison to related studies, the best results achieved with sentence template classification (feature set L and LG) on the ECB+ are comparable to results achieved in related work on the ECB. The approach of Lee et al. (2012) reached 55.9% CoNLL-F [5] on the ECB but on a more difficult task entailing mention extraction. Another study reporting the CoNLL F score was done by Cybulska and Vossen (2013) who reached 69.8% CoNLL F1 on the ECB with a component similarity method but on a simpler - within topic task.

Note that the sentence template approach results were generated on the ECB+ corpus extended with texts capturing an additional layer of event instances from the ECB topics. Consequently, the intra-topic ambiguity in the ECB+ is higher than in the ECB. We did not perform topic clustering before comparing event mentions at the sentence level which makes it the task of the coreference resolver to solve intra- and cross-topic ambiguity between event mentions.

6 Conclusion and Future Work

This paper presents a new approach to event coreference resolution. Instead of performing topic classification before solving coreference between event mentions, as most approaches do, the event template approach compares event mentions at the sentence level. In so doing, the approach focuses on solving coreference between different slots of event descriptions, without relying on topic classification for context disambiguation. As such, this heuristic, which on itself is computationally expensive, can also be used after the primary step of topic classification. Especially in case of data sets with high within topic ambiguity where there are multiple event instances described from the same event type (for instance various instances of a *meeting* event). In the future, we will experiment with combining topic classification with the sentence template approach.

This is the only study which we are aware that employs granularity for event coreference resolution.

[4] In the future we will look into extracting the specific entity types so that they can be used for event coreference resolution, regardless of data annotation.

[5] The CoNLL F measure was used for comparison of competing coreference resolution systems in the CoNLL 2011 task.

For the purpose of this task a new granularity ontology was created. As our method does not employ POS and syntactic role information and no anaphora resolution or topic classification was performed to aid coreference resolution, the results are highly encouraging. In our future work we will look at possibilities of extending the granularity ontology learning granularity levels from corpora to overcome the low coverage limitation following from the usage of a WordNet based taxonomy. We will also augment the ontology to cover the non-human participant slot and experiment with other ways to represent event granularity with features.

Acknowledgments

This work has been carried out within the NewsReader project supported by the EC within the 7th framework programme under grant agreement nr. FP7-ICT-316404. We are grateful for the feedback from the anonymous reviewers. All mistakes are our own.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc., http://nltk.org/book.

Agata Cybulska and Piek Vossen. 2013. Semantic relations between events and their time, locations and participants for event coreference resolution. In *Proceedings of recent advances in natural language processing (RANLP-2013)*.

Agata Cybulska and Piek Vossen. 2014a. Guidelines for ECB+ annotation of events and their coreference. Technical Report NWR-2014-1, VU University Amsterdam.

Agata Cybulska and Piek Vossen. 2014b. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2014)*.

Agata Cybulska and Piek Vossen. 2015. "Bag of events" approach to event coreference resolution. Supervised classification of event templates. In *International Journal of Computational Linguistics and Applications (IJCLA)*.

Andrey Gusev, Nathanael Chambers, Pranav Khaitan, Divye Khilnani, Steven Bethard, and Dan Jurafsky. 2011. Using query patterns to learn the duration of events. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS11)*.

Jerry R. Hobbs. 1985. Granularity. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*.

Blake Stephen Howald and Martha Abramson. 2012. The use of granularity in rhetorical prediction. In *Proceedings of the First Joint Conference on Lexi-cal and Computational Semantics (*SEM)*.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *ANARESOLUTION '97 Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Catharina Maria Keet. 2008. A formal theory of granularity. toward enhancing biological and applied life sciences information systems with granularity. In *Ph.D. thesis, Faculty of Computer Science, Free University of Bozen-Balzano, Italy*.

LDC. 2005. ACE (Automatic Content Extraction) English Annotation Guidelines for Events ver. 5.4.3 2005.07.01. In *Linguistic Data Consortium*.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*.

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

Inderjeet Mani. 1998. A theory of granularity and its application to problems of polysemy and underspecification of meaning. In *In Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference (KR-98)*.

Rutu Mulkar-Mehta, Jerry R. Hobbs, and Eduard Hovy. 2011a. 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*.

Rutu Mulkar-Mehta, Jerry R. Hobbs, and Eduard Hovy. 2011b. Granularity in natural language discourse. In

Proceedings of International Conference on Computational Semantics.

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and douard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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 CoNLL 2011: Shared Task*.

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

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

Morton E. Winston, Roger Chaffin, and Douglas Herrmann. 1987. A taxonomy of part-whole relations. In *Cognitive Science Volume 11, Issue 4, pages 417 - 444*.