International Journal of Computational Linguistics and Applications, Vol. 6, No. 2, 2015, pp. 11–27
Received 25/01/2015, Accepted 25/02/2015, Final 18/11/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

"Bag of Events" Approach to Event Coreference Resolution. Supervised Classification of Event Templates

AGATA CYBULSKA
PIEK VOSSEN
Free University Amsterdam, Amsterdam

ABSTRACT

We propose a new robust two-step approach to cross-textual event coreference resolution on news articles. The approach makes explicit use of event and discourse structure thereby compensating for implications of the Gricean Maxim of quantity. News follows the principle of language economy. Information tends not to be repeated within discourse boarders. This phenomenon poses a challenge for models comparing information about event mentions (and their arguments) on the sentence level. Our approach addresses this challenge by building a knowledge representation per unit of discourse - for present purposes, a document. We collect event information from a single document filling in a "document template" and by that creating a "Bag of Events." We then use supervised Classification to determine if pairs of document templates contain corefering event mentions. Next we solve coreference between event mentions from the same document cluster by means of supervised classification of "sentence templates." The results indicate that the new approach is promising.

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 the news.

It is common practice to use information coming from event arguments for event coreference resolution ([1], [2], [3], [4], [5], [6], [7], [8] among others). The research community seems to agree that event context information regarding time and place of an event as well as information about other participants play an important role in resolution of coreference between event mentions. Even though the contribution coming from event arguments as calculated in some studies does not directly translate into some significant increase of coreference resolution scores. [2] report that features related to event arguments slightly (+2.4% ECM F) improve intra-document event coreference. [7] note a ca. 4% CoNLL F-score improvement of within-topic event coreference resolution based on semantic similarity of event arguments.

Using entities for event coreference resolution is made complicated by the fact that descriptions of events at the sentence level often lack some pieces of information. As pointed out by [1], it could be the case however that a lacking piece of information might be available elsewhere within discourse boarders. News articles can be seen as a form of public discourse [9]. As such, the news follows the Gricean Maxim of quantity [10]. Journalists do not make their contribution more informative than necessary. This means that some information previously communicated within a unit of discourse, will not be mentioned again, unless pragmatically required. This is a challenge for models comparing mentions of events (and their arguments) with one another on the sentence level. One would like to be able to fully make use of information coming from event arguments. Instead of looking at event information available within the same sentence, we propose to take a broader look at event mentions surrounding the event mention in question within a unit of discourse. For the purpose of this study, we consider a document (here a news article) to be our unit of discourse.

This study experiments with an “event template” approach which employs the structure of event descriptions for event coreference resolution. In the proposed heuristic, event mentions

are examined through the perspective of five slots, as annotated in the dataset used in our experiments. The event slots correspond to different elements of event information: an event action (or an event trigger following the ACE terminology [11]) and four types of event arguments: time, location, human and non-human participant slots (see [12], whose ECB+ corpus [13] annotated with event coreference will be used in the experiments). The approach employed in this paper determines coreference between descriptions of events through compatibility of slots of an event template. Figure 1 presents an excerpt from topic 1, text number 7 of the ECB corpus [5]. Consider two event template examples presenting the distribution of event information over the five event slots in the two example sentences (Table 1).

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

Figure 1. Text 7, topic 1, ECB corpus [5]

An event template can be filled from different units of discourse, such as a sentence, a paragraph or an entire document. We propose a two-step classification approach to event coreference resolution. In the first step of the approach, an event template is filled in per document; this is a "document template." By filling in a document template, one creates a "Bag of Events" per document. Bag of Events features are then used in supervised Classification.

This heuristic employs clues coming from discourse structure and namely those implied by discourse boarders. Descriptions of different event mentions occurring within a discourse unit, whether coreferent or related in some other way, unless stated otherwise, tend to share their context. In the example fragment (Figure 1) the first sentence reveals that an actress has entered a rehab facility. From the second sentence the reader finds out where the facility is located and when the actress headed there. It is clear to the reader of the example text fragment from Figure 1

that both event mentions from sentence one and two, happened on Tuesday. Also both sentences mention the same rehab center in Malibu. These observations are crucial for the Bag of Events approach proposed here. As the first step of the approach a document template is filled, accumulating mentions of the five event slots from a document, as exemplified in Table 1. Supervised classifiers determine whether pairs of document templates contain any corefering event mentions. In the second step of the approach coreference is solved between event mentions within document clusters created in step 1. For the purpose of this task again an event template is filled but this time, it is a "sentence template" which gathers event information from the sentence per action mention. Supervised classifiers solve coreference between pairs of event mentions and finally pairs sharing common mentions are chained into coreference clusters.

Table 1. *Sentence and document templates ECB topic 1, text 7, sentences 1 and 2*

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

The main contribution of this work is the new robust Bag of Events approach to event coreference resolution that accounts for the realization of Gricean maxim of quantity in the news by incorporating Bag of Events features. The two-step Classification approach replaces the typically used in coreference resolution topic Classification step with document template Classification that allows for more specific event context disambiguation also within the same topic. Furthermore, the Bag of Events approach implies data representation through a relatively small number of features and yet delivers results comparable to those achieved in related work employing extended feature sets.

We will first delineate the Bag of Events approach to event coreference resolution in section 2. Section 3 reports on the experiments with the new method. We compare the results reached by means of our approach to those from related work in section 4. We conclude in section 5.

2. TWO-STEP BAG OF EVENTS APPROACH

We present a novel two-step approach to cross-textual event coreference resolution on news articles that explicitly employs event and discourse structure to account for implications of Gricean maxim of quantity. The first step in this approach is to build a knowledge representation by filling in an event template per unit of discourse - here a document. We collect all event action, location, time, human and non-human participant mentions from a single document and we fill in a document template (as depicted in Table 1). We then find pairs of document templates containing corefering event mentions by means of supervised Classification. In the second step, we use supervised classifiers to solve coreference between pairs of event mentions within clusters of document templates as determined in step 1. These steps are described in more detail below. Figure 2 depicts the implications of the two-step approach for the training data and Figure 3 for the test set.

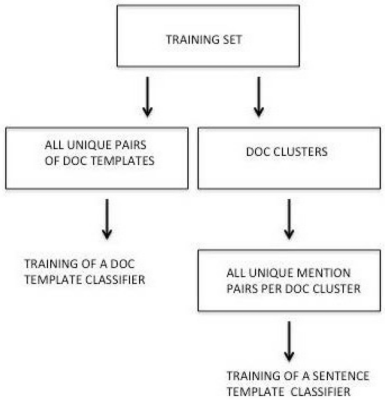

Figure 2. *Training set processing*

Figure 3. *Test set processing*

Step 1. Clustering document templates

The first step in this approach is to fill in a document template. We create a document template by collecting mentions of the five event slots: action, location, time, human and non-human participant from a single document. In a document template there is no distinction made between pieces of event information coming from different sentences of a document and no information is kept about elements being part of different mentions. A document template can be seen as a Bag of Events (and event arguments). The template stores a set of unique lemmas per event slot.

On the training set of the data, we train a pairwise binary classifier to determine whether two document templates share corefering event mentions. This is a supervised learning task in which we determine "compatibility" of two document templates if any two mentions from those templates were annotated in the corpus as coreferent. Let m be an event mention, and *doc* a collection of mentions from a single document template such that fm $\{m_i: 1 \leq i \leq doc_j\}$ where i is the index of a mention and J indexes document templates; doc_j: $1 \leq j \leq DOC$ where *DOC* are all document templates from the corpus. Let ma and mb be mentions from different document templates. "Compatibility" of a pair of document templates (doc$_j$; doc$_j$+1) is determined based on coreference of any mentions (ma_i, mb_i) from a pair of document templates such that:

$$coref\ (\exists ma_i \in doc_j,\ \exists mb_i \in doc_{j+1}) \Rightarrow compatibility\ (doc_j;\ doc_{j+1}).$$

On the training data, we train a binary decision-tree classifier (hereafter DT) to find pairs of document templates containing corefering event mentions.

After all unique pairs of document templates from the test set are classified by means of the DT document template classifier, "compatible" pairs are merged into document clusters based on pair overlap.

Step 2. Clustering sentence templates

The aim of the second step is to solve coreference between event mentions from document clusters which are the output of the Classification task from Step 1. We experiment with a supervised decision tree classifier. This time in the classification task pairs of sentence templates are considered.

A sentence template (e.g. Table 1) is created for every event action mention in the data set. All unique pairs of event mentions (and their sentence templates) are generated within clusters of documents sharing corefering event mentions in the training set. Pairs of sentence templates, that translate into features indicating compatibility across five event template slots, are used to train a decision tree sentence template classifier.

On the test set part of the data; after output clusters of the document template classifier from step 1 are turned to mention pairs (all unique action mention pairs within a document cluster), pairs of sentence templates are classified by means of the DT sentence template classifier. To identify the final equivalence classes of corefering event mentions, within each document cluster, event mentions are grouped into equivalence classes based on corefering pair overlap.

3. EXPERIMENTS

3.1. *Corpus*

For the experiments we used true mentions from the ECB+ corpus [13] which is an extended and re-annotated version of the ECB corpus [5]. 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 resolution 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 2 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).

3.2. *Experimental set up*

The ECB+ texts are available in the XML format. The texts are tokenized, so no sentence segmentation nor tokenization needed to be done. We POS-tagged (for the purpose of proper verb lemmatization) and lemmatized the corpus sentences. For the experiments we used tools from the Natural Language Toolkit ([14], NLTK version 2.0.4): the NLTK's default POS tagger, and WordNet lemmatizer[1] as well as WordNet synset assignment by the NLTK[2]. For machine learning experiments we used scikit-learn [15].

In the experiments, different features were assigned values per event slot (see Table 3). The lemma overlap feature (L) expresses a percentage of overlapping lemmas between two instances of an event slot, if instantiated in the sentence (with the exclusion of stop words). As the relation between an event and involved entities is not annotated in ECB+, frequently one ends up with multiple entity mentions from the same sentence for an action mention. All entity mentions from the sentence are considered in the overlap calculations. There are two features indicating event mentions' location within discourse (D), specifying if two mentions come from the same sentence and the same document. Action similarity (A) was calculated for a pair of active action mentions using the Leacock and Chodorow measure [16]. Per entity slot (location, time, human and non-human participant) we checked if there is any coreference relations annotated in the corpus between entity mentions from the

[1] www.nltk.org/ modules/nltk/stem/wordnet.html
[2] http://nltk.org/ modules/nltk/corpus/reader/wordnet.html

sentence for the two compared event actions; we used cosine similarity to express this feature (E). For all five slots a percentage of synset overlap is calculated (S). In case of document templates features referring to active action mentions were disregarded, instead only action mentions from a document were considered. All feature values were rounded to the first decimal point.

Table 2. *ECB+ statistics*

ECB+	#
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 3. *Features grouped into four categories: L-Lemma based, A-Action similarity, D-location within Discourse, E-Entity coreference and S-Synset based*

Event Slot	**Mentions**	**Feature Kind**	**Explanation**
Action	Active mentions	Lemma overlap (L) Synset overlap (S) Action similarity (A) Discourse location (D) - document - sentence	Numeric feature: overlap %. Numeric: overlap %. Numeric: [16]. Binary: - the same document or not. - the same sentence or not.
	Sent. or doc. mentions	Lemma overlap (L) Synset overlap (S)	Numeric: overlap %. Numeric: overlap %.
Location	Sent. or doc mentions	Lemma overlap (L) Entity coreference (E) Synset overlap (S)	Numeric: overlap %. Numeric: cosine similarity. Numeric: overlap %.
Time	Sent. or doc mentions	Lemma overlap (L) Entity coreference (E) Synset overlap (S)	Numeric: overlap % Numeric: cosine similarity. Numeric: overlap %.
Human Participant	Sent. or doc mentions	Lemma overlap (L) Entity coreference (E) Synset overlap (S)	Numeric: overlap %. Numeric: cosine similarity. Numeric: overlap %.
Non-Human Participant	Sent. or doc mentions	Lemma overlap (L) Entity coreference (E) Synset overlap (S)	Numeric: overlap %. Numeric: cosine similarity. Numeric: overlap %.

We experimented with a few feature sets, considering per event slot lemma features only (L), or combining them with other features described in Table 3. Before fed to a classifier, missing values were imputed (no normalization was needed for the scikit-learn DT algorithm). All classifiers were trained on an unbalanced number of pairs of document or sentence templates from the training set. We used grid search with ten fold cross-validation to optimize the hyper-parameters (maximum depth, criterion, minimum samples leafs and split) of the decision-tree algorithm.

3.3. *Baseline*

We will consider two baselines: a singleton baseline and a rule-based lemma match baseline. The singleton baseline considers 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. The rule-based lemma baseline generates event mention coreference clusters based on full overlap between lemma or lemmas of compared event triggers (action slot) from the test set.

Table 5 presents baselines' results in terms of recall (R), precision (P) and F-score (F) by employing the coreference resolution evaluation metrics: MUC [17], B3 [18], CEAF [19], BLANC [20], and CoNLL F1 [21]. When discussing event coreference scores must be noted that some of the commonly used metrics depend on the evaluation data set, with scores going up or down with the number of singleton items in the data [20]. Our singleton baseline gives us zero scores in MUC, which is understandable 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 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. The lemma baseline reaches 62% CoNLL F1. A baseline only considering event triggers, will allow for an interesting comparison with our event template approach, employing event argument features.

3.4. *Evaluation*

Table 4 evaluates the final clusters of corefering event action mentions produced in the experiments by means of the DT algorithm when employing different features. The best coreference evaluation scores with the highest CoNLL F-score of 73% and BLANC F of 72% were reached by the combination of the document template classifier using feature set L across event slots and the sentence template classifier when employing features LDES (see Table 3 for feature de-scription). Adding action similarity (A) on top of LDES features does not make any difference on decision tree classifiers with a maximum depth of 5. Our best CoNLL F-score of 73% is an 11% improvement over the strong rule based event trigger lemma baseline, and a 34% increase over the singleton baseline.

Table 4. *Bag of Events approach to event coreference resolution, evaluated in MUC, B3, mention-based CEAF, BLANC and CoNLL F on the ECB+ corpus*

Step1			Step2			MUC			B3			CEAF	BLANC			CoNLL
Alg	Slot Nr	Feats	Alg	Slot Nr	Feats	R	P	F	R	P	F	F	R	P	F	F
-	-	-	DT	5	L	61	76	68	66	79	72	61	67	69	68	70
DT	5	L	DT	5	L	71	75	73	71	77	74	64	71	71	71	73
DT	5	L	DT	5	LDES	71	75	73	71	78	74	64	72	71	72	73
DT	2	L	DT	2	LDES	76	70	73	74	68	71	61	74	68	70	70
DT	5	L	DT	5	LADES	71	75	73	71	78	74	64	72	71	72	73

To quantify the contribution of document templates, we contrast the results of the two-step Bag of Events approach with scores achieved when skipping step 1 that is without the initial Classification of document templates. The results obtained with sentence template Classification only give us some insights into the impact of the document template Classification step. Note that the sentence template Classification without preliminary document template clustering is computationally much more expensive than the two-step template approach, which ultimately takes into account significantly less item pairs owing to the initial document template clustering. In the one-step approach the DT

sentence template classifier using lemma features (L), when trained on an unbalanced training set, reaches 70% CoNLL F. This is 8% better than the strong lemma baseline disregarding event arguments, but only 3% less than the two-step Bag of Events approach with the two classifiers trained on lemma features (L). The reason for the relatively small improvement by the document template classification step could arise from the fact that in the ECB+ corpus few sentences are annotated per text. 1840 sentences are annotated in 982 corpus texts, i.e. 1.87 sentence per text. We expect that the impact of document templates would be bigger if more event descriptions from a discourse unit were taken into account than only the ground truth mentions.

Table 5. *Baseline results: Singleton baseline and lemma match of event triggers evaluated in MUC, B3, mention-based CEAF, BLANC and CoNLL F*

Baseline	MUC			B3			CEAF	BLANC			CoNLL
	R	P	F	R	P	F	F	R	P	F	F
Singleton Baseline	0	0	0	45	100	62	45	50	50	50	39
Action Lemma Baseline	71	60	65	68	58	63	51	65	62	63	62

We ran an additional experiment with the four entity types bundled into one entity slot. Locations, times, human and non-human participants were combined into a cumulative entity slot resulting in a simplified two-slot template. When using two-slot templates for both, document and sentence classification on the ECB+ 70% CoNLL F score was reached. This is 3% less than with five-slot templates.

4. RELATED WORK

To the best of our knowledge, the only related study using clues coming from discourse structure for event coreference resolution was done by [1] who perform coreference merging between event template structures. Both approaches determine event compatibility within a discourse representation but we achieve that in a different way - with a much more restricted template

(five slots only) which facilitates merging of all event and entity mentions from a text as the starting point. [1] consider discourse events and entities for event coreference resolution while operating on the level of mentions.

Some of the metrics used to score event coreference resolution are dependent on the number of singleton events in the evaluation data set [20]. Hence 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. To the best of our knowledge, no baseline has been set yet for event coreference resolution on the ECB+ corpus. Accordingly, in Table 6 we will also look at results achieved on the ECB corpus which is a subset of ECB+, and so the closest to the data set used in our experiments but capturing less ambiguity of the annotated event types [13]. We will focus on the CoNLL F measure that was used for comparison of competing coreference resolution systems in the CoNLL 2011 shared task.

Table 6. *The Bag of Events (BOE) approaches evaluated on ECB and ECB+ in MUC, B3, entity-based CEAF, BLANC and CoNLL-F in comparison with related studies. Note that the BOE approaches use gold and related studies system mentions*

Approach	Data	Model	MUC			B3			CEAF entity	BLANC			Co-NLL
			R	P	F	R	P	F	F	R	P	F	F
B&H	ECB annotation [5]	HDp	52	90	66	69	96	80	71	NA	NA	NA	NA
LEE	ECB annotation [6]	LR	63	63	63	63	74	68	34	68	79	72	55
BOE-2	ECB annotation [13]	DT+DT	65	59	62	77	75	76	72	66	70	67	70
BOE-5	ECB annotation [13]	DT+DT	64	52	57	76	68	72	68	65	66	65	66
BOE-2	ECB+ annotation [13]	DT+DT	76	70	73	74	68	71	67	74	68	70	70
BOE-5	ECB+ annotation [13]	DT+DT	71	75	73	71	78	74	71	72	71	72	**73**

The best results of 73% CoNLL F were achieved on the ECB+ by the Bag of Events approach using five slot event templates (*BOE-5* in Table 6). When using two-slot templates we get 3% less CoNLL F on ECB+. For the sake of comparison, we run an additional experiment on the ECB part of the corpus (annotation by [13]). The ECB was used in related work although with different versions of annotation so not entirely comparable. We

run two tests, one with the simplified templates considering two slots only: action and entity slot (as annotated in the ECB by [6]) and one with five-slot templates. The two slot Bag of Events (*BOE-2*) on the ECB part of the corpus reached comparable results to related works: 70% CoNLL F, while the five-slot template experiment (*BOE-5*) results in 66% CoNLL F. The approach of [6] (in Table 6 *LEE*) using linear regression (in Table 6 *LR*) reached 55% CoNLL F although on a much more difficult task entailing event extraction as well. The component similarity method of [7] resulted in 70% CoNLL F but on a simpler within topic task (not considered in Table 6). *B&H* in Table 6 refers to the approach of [5] using hierarchical Dirichlet process (*HDp*); for this study no CoNLL F was reported. In the *BOE* experiments reported in Table 6, during step 1 only lemma features L were used and for sentence template Classification (step 2) LDES features were employed. In all tests with the Bag of Events approach, ground truth mentions were used.

5. CONCLUSION

This paper presents a two-step Bag of Events approach to event coreference resolution. Instead of performing topic Classification before solving coreference between event mentions, as is done in most studies, this two-step approach first compares document templates created per discourse unit. Only after does it compare single event mentions and their arguments. In contrast to a heuristic using a topic classifier, that might have problems distinguishing between multiple instances of the same event type, the Bag of Events approach facilitates context disambiguation between event mentions from different discourse units. Grouping events depending on compatibility of event context (time, place and participants) on the discourse level, allows one to take advantage of event context information, which is mentioned only once per unit of discourse and consequently is not always available on the sentence level. From the perspective of performance, the robust Bag of Events approach using a small feature set also significantly restricts the number of compared items. Therefore, it has much lower memory requirements than a

pairwise approach operating on the mention level. Given that this approach does not consider any syntactic features and that the evaluation data set is only annotated with 1.8 sentences per text, the evaluation results are highly encouraging. Future research will be dedicated to experimenting with the Bag of Events approach on event slot mentions extracted by the system to demonstrate conclusively the validity of the approach.

REFERENCES

1. Humphreys, K., Gaizauskas, R. & Azzam, S. 1997. Event coreference for information extraction. In *ANARESOLUTION '97 proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.
2. Chen, Z. & Ji, H. 2009. Event coreference resolution: Feature impact and evaluation. In proceedings of *Events in Emerging Text Types (eETTs) Workshop*.
3. Chen, Z. & Ji, H. 2009. Graph-based event coreference resolution. In *TextGraphs-4 Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing* (pp. 54-57).
4. Chen, B., Su, J., Pan, S. J., Tan, C. L. 2011. A unified event coreference resolution by integrating multiple resolvers. In proceedings of the *5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand.
5. Bejan, C. A. & Harabagiu, S. 2010. Unsupervised event coreference resolution with rich linguistic features. In proceedings of the *48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.
6. Lee, H., Recasens, M., Chang, A., Surdeanu, M. & Jurafsky, D. 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)*.
7. Cybulska, A. & Vossen, P. 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)*.
8. Liu, Z., Araki, J., Hovy, E. & Mitamura, T. 2014. Supervised within-document event coreference using information propagation. In proceedings of the *International Conference on Language Resources and Evaluation (LREC 2014)*.

9. van Dijk, T.A. 1988. News As Discourse. Routledge.
10. Grice, P. 1975. Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (pp. 41-58). New York: Academic Press.
11. LDC. 2005. ACE (Automatic Content Extraction) English Annotation Guidelines for Events ver. 5.4.3 2005.07.01. In *Linguistic Data Consortium*.
12. Cybulska, A. & Vossen, P. 2014. Guidelines for ECB+ annotation of events and their coreference. *Technical Report NWR-2014-1*, VU University Amsterdam.
13. Cybulska, A. & Vossen, P. 2014. 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)*.
14. Bird, S., Klein, E. & Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc., http://nltk.org/book.
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
16. Leacock, C. & Chodorow, M. 1998. Combining local context with wordnet similarity for word sense identification. In *WordNet: A Lexical Reference System and its Application*.
17. Vilain, M., Burger, J., Aberdeen, J., Connolly, D. & Hirschman, L. 1995. A model theoretic coreference scoring scheme. In *Proceedings of MUC-6*.
18. Bagga, A. & Baldwin, B. 1998. Algorithms for scoring coreference chains. In proceedings of the *International Conference on Language Resources and Evaluation (LREC)*.
19. Luo, X. 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)*.
20. Recasens, M. & Hovy, E. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17/4, 485-510.
21. Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M.,Weischedel, R. & Xue, N. 2011. Conll2011 shared task: Modeling unrestricted coreference in ontonotes. In proceedings of *CoNLL 2011: Shared Task*.

ACKNOWLEDGMENTS

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

AGATA CYBULSKA
FREE UNIVERSITY AMSTERDAM
DE BOELELAAN 1105 1081HV AMSTERDAM
E-MAIL: <A.K.CYBULSKA@VU.NL>

PIEK VOSSEN
FREE UNIVERSITY AMSTERDAM
DE BOELELAAN 1105 1081HV AMSTERDAM
E-MAIL: <PIEK.VOSSEN@VU.NL>