Highly Multilingual Coreference Resolution Exploiting a Mature Entity Repository

Josef Steinberger, **Jenya Belyaeva**, **Jonathan Crawley**, **Leonida Della-Rocca**,
Mohamed Ebrahim, **Maud Ehrmann**, **Mijail Kabadjov**,
Ralf Steinberger and **Erik van der Goot**
EC Joint Research Centre
21027, Ispra (VA), Italy
name.surname@jrc.ec.europa.eu

Abstract

In this paper we present an approach to large-scale coreference resolution for an ample set of human languages, with a particular emphasis on time performance and precision. One of the distinctive features of our approach is the use of a mature multilingual named entity repository (persons and organizations) gradually compiled over the past few years. Our experiments show promising results – an overall precision of 94% tested on seven different languages. We also present an extrinsic evaluation on seven languages in the context of summarization where we gauge the contribution of the coreference resolver towards the end summarization performance.

1 Introduction

Recent work on coreference resolution has been largely dominated by machine learning approaches and predominantly for the English language (Ng and Cardie, 2002; Ponzetto and Strube, 2006; Luo, 2007). This is in great part due to the availability of annotated corpora such as MUC-6/7 (Hirschman, 1998), ACE-2/3/4/5 (NIST, 2004), GNOME (Poesio et al., 2004) and large-scale crowdsourcing efforts like Phrase Detectives.[1]

One of the big advantages of machine learning approaches is that they are reasonably easy to reproduce given that the set of input features are documented well, since there are many good open-source platforms for machine learning (e.g., WEKA[2]) and machine-learning-based coreference (e.g., BART[3] (Versley et al., 2008)).

However, intrinsic evaluations can pose problems. As pointed out by (Stoyanov et al., 2009) there is too much variation in reported results across data sets to be able to draw robust conclusions on the state-of-the-art in the area for which they proposed a method for reporting results on a data set that makes it easier to predict performance on other data sets (by breaking down results into names, types of pronouns, nominals etc.). Also, intrinsic evaluations can be highly sensitive to preprocessing (Mitkov, 2002).

There is agreement in the community on the level of resolution difficulty on major types of coreferential expressions. For instance, proper names are considered to be the easiest to resolve, followed by pronouns, in turn followed by common nouns. One of the main reasons why common noun coreference is challenging is because they often share little or no surface linguistic features with their antecedents and require world or encyclopedic knowledge for their resolution (see (Kabadjov, 2007) for a study for English). For instance, Ponzetto and Strube (2006) proposed to use WordNet and Wikipedia to address the problem of bringing in world and/or encyclopedic knowledge into their system for coreference resolution in English reporting improvements for common noun resolution.

In this work we address two important remaining gaps in coreference resolution. Firstly, we are interested in highly multilingual coreference. Secondly, we address the problem of common noun coreference by exploiting a large lexical resource, the named entity database, compiled over the past few years by automatically extracting names from hundreds of thousands of online news articles in twenty languages (and subsequently cleaning the most frequent names by a human moderator). The coreference resolver we present is designed to work as part of the Europe Media Monitor (EMM) system[4] for online news analysis and aggregation.

[1] http://www.phrasedetectives.org.

[2] http://www.cs.waikato.ac.nz/ml/weka/.

[3] http://www.bart-coref.org/.

[4] http://emm.newsbrief.eu/overview.html

Proceedings of Recent Advances in Natural Language Processing, pages 254–260,
Hissar, Bulgaria, 12-14 September 2011.

In order to evaluate the effectiveness of our approach we carry out two separate evaluations: one intrinsic and one extrinsic in the context of summarization.

The rest of the paper is organized as follows: in the next section (2) we describe our named entity database which is the backbone of our approach; in 3, we present our approach to coreference followed by a discussion of experimental results in 4. Then, in 5 we briefly survey related work on coreference resolution and finally conclude and give pointers to future work.

2 The Multilingual Named Entity Database

The historical repository of EMM's person and organization titles is a by-product of the Named Entity Recognition (NER) process, which has been applied daily to tens of thousands of multilingual news articles per day since 2004. Titles are parts of the name recognition patterns, and each time a name is found, EMM keeps track of the titles found next to the name. The result is a large multilingual repository of titles and other attributes about names. In this section, we thus try to give an overview of the NER process and hence information about the title repository.

EMM's NER is performed by applying language-independent recognition patterns to text. The hand-written language-independent recognition patterns use slots to make reference to various language-specific lists of words, phrases and regular expressions. By doing this, the system is modular and a new language can simply be plugged in by adding the language-specific parameter file, containing the relevant word lists for each slot. Pouliquen and R. Steinberger (2009) describe the types of slots and list a number of patterns. A typical and simple pattern is the one that requires that uppercase words adjacent to any title are likely to be person or organization names (e.g., *President* `Upper Upper`). As the strings indicating that neighboring uppercase words in a name are not necessarily titles, we refer to them more generally as Trigger Words. The trigger word list of elements thus contains conventional titles (e.g., *Dr.*, *Mr.*, *President*), professions and occupations (e.g., *spokeswoman*, *artist*, *playboy*, *tennis player*), roles inside teams (*secretary*, *defense player*, *short-stop*), adjectives referring to countries, regions, locations, ethnic groups or religions (e.g., *Iraqi*, *Latin-American*, *Parisian*, *Berber*, *Catholic*), and a variety of other strings that may indicate that the adjacent uppercase words are a person (e.g., *XX-year-old*, *has declared*, *deceased*). These lists are mostly produced using empirical methods or machine learning, but they are always manually verified. The rules are partially cascaded and allow for large combinations of trigger words, e.g., to recognize the uppercase words in the following apposition construction as a name: `Upper Upper`, *former 56-year-old Afghan Foreign Minister*.

As the patterns exist and are applied to twenty languages, the list of trigger words contains words in all these languages. Some of these trigger words are not suitable so we remove them from the lists. Age expressions such as *XX-year-old* or verbal phrases such as *has declared* were thus manually removed.

Patterns to recognize organizations have different shapes and the trigger words are usually part of the organization name (e.g., *Bank* and *Club* in *Chartered Bank* or *Motor Sport Club*). These typical organization name parts are also used for the co-reference resolution task.

3 Coreference Algorithm

3.1 System Architecture

The coreference resolution module is built for inclusion in a larger pipeline architecture, where an input text document undergoes several processing phases during which the source is augmented with layers of meta data such as named entities. The data interchange format between processing phases is RSS, a light-weight type of XML typically used by on-line news providers.

3.2 Lookup of Known Named Entities

Known entities are entities that have been found in at least five different news clusters in the past in the EMM system. For all known entities morphological or other spelling variants are automatically generated according to hand-written rules. For example, for *Angela Merkel*, the genitive version *Merkels* will be pre-generated and recognized, and Arabic names using the infix *al* will be pre-generated with and without *al*, as well as with and without linking hyphens (*Moussab al-Zarqawi*, *Moussab al Zarqawi*, *Moussab Zarqawi*). For the actual lookup, a finite state tool that allows patterns and partial case sensitivity is used, employing entity information that has been gathered over a number of years from the EMM production system to recognize known entities within the text (currently, there are over 1.2 million distinct entities in the named entity repository). The RSS is then marked up with additional meta information about the entities found (see (Crawley and Wagner, 2010) for more details).

3.3 Entity Guessing

As we are interested in grounding name references to real-life entities and we thus need to disambiguate between people having the same surname (or first name), we only look for entities consisting of at least two name parts.

The entity guessing comprises two parts, the first is a parallel lexical tokenization of the text, using classifying tokenizers, gazetteers, pattern matchers and simple tokenizers as well as any previously defined entities from further up the processing chain. The second part is a sequence of finite state grammars that pick and choose appropriate tokens for a given rule from the parallel token streams passing the output on to the next grammar in the sequence building ever more complex constructs and disambiguating on the way.

3.4 Merging of NE Variants

The entity normalization takes place once the entities have been discovered and is used as a means of merging entities with newly found aliases, such as when an existing entity is written in a script we have not seen it in before or has been slightly misspelt. This is done by transliterating the name from any unicode range into the Latin unicode range using a statistical matrix for ngram substitutions. Some normalization may be performed and vowels are removed to create a consonant signature which is then used to perform a lookup for the most likely candidates with the list of known entities. This is to reduce the number of values for eventual comparison using a string similarity metric. The closest match is then selected and, if within a fine-grained tolerance, the value is assigned as a new alias. Otherwise it is assumed a new entity and assigned a new id.

3.5 Coreference Resolver

When an RSS file reaches the coreference resolution module, it already contains the list of known and guessed entities. The resolution is run only over the known entities. The resolver module does the following for each article:

1. Loads all known and guessed entities
2. For each known entity it searches the resources for its possible references (titles from the entity-title table, name parts directly from the entity mention).
3. The reference-entity map is created; it associates each possible reference (step 2) to a known entity.[5]
4. The matcher component finds all possible mentions of any entity (i.e., name parts[6], titles) in the text.[7]
5. The resolver links mentions (step 4) to entities using the reference-entity map, given that the following conditions are met:
   (a) The entity has been already introduced.[8]
   (b) The entity reference is not a constituent of a known or guessed entity mention (or their title).
6. The resolved mentions are merged in order to create a non-overlapping sequence of entity mentions with the following rules:
   (a) If the mention is part of a longer mention leave only the longer one (e.g., 'former US president' would outweigh 'president').
   (b) If the mentions are next to each other and they are assigned to the same entity they are concatenated.
   (c) If the mentions are next to each other and they are assigned to a different entity a name part will outweigh a title (probably an incorrect title).
   (d) Otherwise consider only the latter mention.

4 Evaluation

We carry out a precision-focused intrinsic evaluation over EMM data and an extrinsic evaluation in the context of summarization where we measure the contribution of coreference towards summarization performance. We describe each in turn below.

4.1 Intrinsic Evaluation: EMM Data

In order to evaluate our coreference system we compiled a corpus of news articles in seven different languages: English, German, Italian, Spanish, French, Russian and Arabic, thus, covering a

[5] Ambiguous references are ignored (e.g., title 'president' is not considered as a coreference candidate in the case of an article in which two entities carry the title 'president').

[6] We are also aware of names with infixes like 'de la Vega'.

[7] Because of efficiency reasons it uses lists of all possible name parts and titles, not only those found in the article – the resources are loaded during the matcher's initialization.

[8] The candidate mention appears after the first mention of the entity identified by the name recognition module.

Table 1: Corpus statistics.

Language	News articles	Words	Words per art.
English	149	56891	382
German	45	18213	405
Italian	117	14082	120
Spanish	94	18772	200
French	96	35046	365
Russian	149	24435	164
Arabic	67	24400	364
Overall	717	191839	268

diverse set of language family branches as are Germanic, Romance, Slavic and Semitic.[9]

Statistics about the corpus are shown in table 1. Overall, we gathered 717 news articles containing almost 200k words.

4.1.1 Corpus and Quick Annotation

We ran each news article through the EMM pipeline. After that we asked native speakers of the seven languages to go over the news articles and mark whether each highlighted mention points to the correct entity or not, whereby measuring precision.[10] A highlighted mention could be one of three things: a known named entity recognized by the named entity disambiguation system, a mention of an entity guessed by the named entity guesser, or a mention recognized and attached to a coreference chain by the coreference resolver. The human subjects marked each entity mention via a simple HTML interface.

4.1.2 Results and Discussion

We present separate performance results for named entity disambiguation (table 2) and for coreference resolution (table 3). In both cases we report precision.

Overall, the named entity disambiguation precision was high; 95% of the 2631 named entities recognized by the system were correct (see table 2). The recognition precision of person names in Arabic was the lowest, 81.7%. We discuss the possible reasons for that in our detailed error analysis below. The type of entities entailed by the category ‘Others’ is mostly mentions to organizations, but also some other prominent named entities such as events (e.g., Woodstock Festival).

Table 2: Quality of named entity recognition in the analyzed languages. Values correspond to: Precision (Correct/Recognized).

Language	Persons	Others	All
English	97.0% (419/432)	89.5% (256/286)	94.0% (675/718)
German	97.5% (230/236)	100.0% (46/46)	97.9% (276/282)
Italian	92.1% (151/164)	100.0% (76/76)	94.6% (227/240)
Spanish	95.7% (180/188)	96.0% (72/75)	95.8% (252/263)
French	98.4% (432/439)	97.2% (278/286)	97.9% (710/725)
Russian	97.7% (130/133)	100.0% (35/35)	98.2% (165/168)
Arabic	81.7% (125/153)	100.0% (82/82)	88.1% (207/235)
Overall	95.5% (1667/1745)	95.4% (845/886)	95.5% (2512/2631)

Table 3: Quality of coreference resolution. Values correspond to: Precision (Correct/Recognized).

Language	Person name parts	Person titles	Organiz. head nouns	All
English	99.2% 237/239	72.7% 40/55	94.4% 34/36	94.2% 311/330
German	99.0% 104/105	86.7% 13/15	100.0% 1/1	97.5% 118/121
Italian	94.1% 16/17	75.0% 9/12	100.0% 1/1	86.8% 26/30
Spanish	100.0% 41/41	72.7% 16/22	100.0% 4/4	91.0% 61/67
French	98.1% 51/52	61.2% 52/85	13.3% 2/15	69.1% 105/152
Russian	100.0% 45/45	100.0% 7/7	– 0/0	100.0% 52/52
Arabic	92.9% 92/99	100.0% 2/2	40.0% 2/5	90.6% 96/106
Overall	98.0% 586/598	70.2% 139/198	71.0% 44/62	89.6% 769/858

We present the coreference performance in three distinct categories: *person name parts*, *person titles* and *organization head nouns* (see table 3).

Not surprisingly, the overall coreference resolution of proper names yields high precision (98%), since resolution difficulty increases as folows: proper names pronouns common noun phrases, in particular definite descriptions. Perhaps more notably, these results provide evidence that this is also the case across languages, with Arabic being lowest with 92.9%.

What is more significant, however, is the performance on *person titles*, which entail mostly refer-

[9] In principle, since the coreference method we propose builds on the named entity repository (2), it can be straightforwardly applied to all the languages covered by the repository (currently 20).

[10] As pointed out earlier, we are interested in precision and not in recall, since the large volume of news articles passing through the EMM pipeline makes up for potential loss in recall.

Table 4: Types of errors.

Type of error	Person name parts	Person titles	Organiz. head nouns	All
Indefinite NP		18	13	32
Res. sparseness		11	3	14
Different POS		18	1	20
Error propag.	9			9
Other	3	12	1	16
Overall	12	59	18	89

ences by means of definite descriptions not sharing a head noun with the antecedent, where the system surpasses the 70% threshold (with the exception of French with 61.2%). It is worth pointing out that these are largely regarded as among the most challenging to resolve, mainly because their resolution requires real-world knowledge.

It should be noted also that our system is an end-to-end system, whose input is free text akin to (Mitkov, 2002; Kabadjov, 2007).

In what follows we discuss several representative examples.

Arabic. In the following example the system recognizes بَابَا (*Pope*) as the correct reference to the preceding recognized person بنديكتوس السَادس عشر (*Benedikt XVI*), because our resources capture that *Pope* is one of the titles of *Benedikt XVI* (بنديكتوس السَادس عشر, 'Benedikt XVI' بَابَا, 'Pope'):

(1) يعتزم بَابَا الفَاتيكَان **بنديكتوس السَادس عشر** القيَام بزيَاره الَى كنيس يهودي في العَاصمة الَايطَالية رومَا في ثَاني زيَارة من نوعهَا في تَاريخ الكنيسة الكَاثولكية. وتَاتي زيَارة ا **البَابَا** في وقت .... الحرب العَالمية الثَانية.

English. And here is a similar example in English:

(2) Bruce, who has until 31 December to respond to the FA's request, had asked [Andre Mariner] to look at Turner's red card again... "I hope [the referee] looks at it again. I doubt it, though."

Russian. And finally an example in Russian (, 'Mahmoud Ahmadinejad' , 'leader'):

(3)

4.1.3 Detailed Error Analysis

In this section we discuss the most prominent types of errors and give illustrative examples for Arabic and French.[11] We adopt a precision-focused error analysis.

Precision-focused analysis of errors. We have grouped system errors into five major categories (see table 4): *indefinite noun phrases* (the system wrongly links an indefinite noun phrase to an antecedent), *resource sparseness* (errors due to incomplete database of names and/or titles), *different part-of-speech* (the system assumes a wrong part-of-speech, e.g., *official* as adjective or noun), *error propagation* (errors at the named entity lookup stage propagate on to the coreference resolution) and a general category *Other* for all the remaining errors. To illustrate these error types, we give a few representative examples next.

Arabic. While working on Arabic articles we were faced with some difficulties related to issues of ambiguity, propagation of errors from the NER module and a relative lack of resources compared to other languages. Ambiguity of Arabic person and organization names is mainly due to the relatively high polysemy of Arabic words, the widespread omission of diacritic vowels in written text and the lack of capitalization in the Arabic writing system. For example, some of the very common person names in Arabic like رمضَان *Ramdan* , شعبَان *Shaban* and رجب *Ragab* also stand for month names, so if we have an Entity called محمد رمضَان *Mohamed Ramdan* and at a later distance in text the word رمضَان *Ramdan*, it is difficult to decide if this is a reference to the previous entity or if it is the name of a month. Moreover, the lack of diacritic vowels increases the number for possible readings for a given word, if we have for example the name سَيد عمر *Sayad Amr* and the name part عمر *Amr* in a non vocalized text, the word عمر *Amr* could have four different meanings,

[11] We left out examples for other languages due to space contraints.

whereas if we had the word in the vocalized form عُمَر *Umar*, the only possible meaning would be that of a proper name. A different kind of ambiguity results from the fact that in most Arabic countries there is no real distinction between first and last names. So, the reference to a person's full name could be done by any of the parts of the name, that is, usually in news articles references to "Saddam Hussein" would use the first part of his name, whereas references to "Muhammad Husni Mubarak" would use the third part of the name.

French. There were several errors due to incorrect recognition of named entity boundaries (i.e., error propagation). For instance, in the following example (example 4), the reference to *Ligue 2* has been wrongly recognized as *Ligue* and subsequently identified as coreferential with *Ligue 1*:

(4) Neuf des dix matches de cette 20e journée de [Ligue 1] sont programmés ce soir à 21h, avec notamment un intéressant Lille-PSG. En bas de tableau, le match de la peur oppose Grenoble, quasiment assuré de descendre en *[Ligue]* 2, à Saint-Etienne, 18e et premier relégable.

4.2 Extrinsic Evaluation via Summarization: Project-Syndicate Data

Kabadjov (2007) argued that Summarization is a suitable task for evaluating extrinsically coreference resolution systems. Here, we take on their proposal and in this section we discuss experiments with an LSA-based summarizer integrated with the coreference resolver described above on a publicly available corpus[12] for evaluating multi-document multilingual[13] summarization systems (Turchi et al., 2010).[14]

Our approach for integrating a coreference resolver into an LSA-based summarization system draws on the method put forward by (Steinberger et al., 2007). The intuition behind this choice is that in addition to capturing pure lexical co-occurrence the extended system is also capable of capturing entity co-occurrence which takes the summarization process to a more semantically-aware level.

4.2.1 Experimental Results

The experimental results are presented in table 5. Each summary score is computed by first calculating the intersection of sentences selected by the summarizer with those selected by at least two annotators divided by the number of sentences in the system summary.[15]

The first thing we observe is that overall (see bottom part of table 5) for target summaries of size three sentences or smaller incorporating cross-document coreference works better than the baseline LSA case and both perform better than two baseline summarizers: one selecting the first sentence of each document in the cluster (labeled 'Lead' in table 5) and another one selecting random sentences (labeled 'Random'). One possible reason for that is that by adopting a more semantically-aware representation the summarization machinery is able to produce succinct summaries of better quality than the LSA-only method, but as soon as the summarization compression rate is relaxed the benefit of including entities becomes less visible (and even in some cases yields worse results).

The variation in summarization performance across languages can be in part explained by the inconsistent performance of the coreference resolver due to lack of or noisy resources for the languages. For instance, for languages like English and German we have good coreference resolution performance which also translates into decent summarization performance, whereas for Czech the performance is notably lower.

5 Related work

Representatives of machine learning work on coreference are (Ng and Cardie, 2002; Luo, 2007) for supervised learning and (Haghighi and Klein, 2007) for unsupervised.

In more recent work, (Stoyanov et al., 2009) provides a comprehensive discussion of the state of the art coupled with extensive experiments on the standard corpora for English: MUC-6, MUC-7, ACE-2, ACE-3, ACE-4 and ACE-5. Recasens and Hovy (2010) explore the impact on coreference resolution performance by varying several prominent contextual factors; they measure performance across corpora, languages, annotation schemes and preprocessing. However, their set of languages consisted of English and Spanish only.

The most closely related experiment to ours is that of the SemEval-2010 task 1 (Recasens et al., 2010), which covered coreference evaluation on six languages.

[12] This is different from the dataset used for the intrinsic evaluation.

[13] Seven languages: English, French, German, Spanish, Russian, Arabic and Czech.

[14] Data publicly available for download at: `http://langtech.jrc.ec.europa.eu/JRC_Resources.html`.

[15] For a discussion on how this evaluation metric compares with ROUGE see (Turchi et al., 2010).

Table 5: Summarization Results.

Summarizers	Summary Size (number of sentences)					
	1	3	5	10	15	20
	English					
LSA+Coref	1.0	.67	.6	.6	.5	.43
LSA	0	.67	.6	.6	.47	.45
	French					
LSA+Coref	.5	.67	.6	.55	.47	.43
LSA	0	.5	.6	.45	.47	.4
	German					
LSA+Coref	1.0	.83	.7	.55	.47	.35
LSA	.5	.5	.7	.55	.43	.38
	Spanish					
LSA+Coref	1.0	.83	.7	.45	.37	.4
LSA	.5	.67	.5	.5	.37	.43
	Russian					
LSA+Coref	1.0	.67	.6	.65	.53	.6
LSA	1.0	.67	.6	.5	.57	.6
	Arabic					
LSA+Coref	0	.5	.7	.55	.47	.5
LSA	.5	.67	.5	.6	.53	.53
	Czech					
LSA+Coref	0	.67	.6	.5	.43	.48
LSA	.5	.67	.7	.7	.53	.48
	Overall					
LSA+Coref	.64	.69	.64	.55	.46	.45
LSA	.43	.62	.6	.56	.48	.46
Lead	-	-	.3	.25	.26	.25
Random	.22	.22	.22	.22	.22	.22

6 Conclusion

In this paper we presented an approach to large-scale coreference resolution for a broad spectrum of human languages with precision and efficiency in mind. The backbone of our algorithm is a mature multilingual named entity database semi-automatically compiled over the past few years.

We reported an overall precision of 94% tested on seven different languages and presented a detailed error analysis with illustrative examples from our corpus.

We performed an extrinsic evaluation on seven languages in the context of the task of summarization. We concluded that producing short informative summaries (from one to three sentences) is better achieved by bringing in cross-document coreference than without it.

In future work, we intend to carry out a comprehensive extrinsic evaluations in the context of end-goal tasks like Sentiment Analysis and Quotation extraction. We also plan to perform an additional intrinsic evaluation on the SemEval'10 corpus.

References

J.B. Crawley and G. Wagner. 2010. Desktop text mining for law enforcement. In *Proceedings of IEEE ISI*, pages 138–140.

A. Haghighi and D. Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of ACL*, pages 848–855.

L. Hirschman. 1998. MUC-7 coreference task definition, version 3.0. In *Proceedings of MUC*. NIST.

M. Kabadjov. 2007. *A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Recognition*. Ph.D. thesis, Department of Computer Science, University of Essex, December.

X. Luo. 2007. Coreference or not: A twin model for coreference resolution. In *Proceedings of NAACL*.

R. Mitkov. 2002. *Anaphora Resolution*. Longman.

V. Ng and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of ACL*.

NIST. 2004. The ace evaluation plan.

M. Poesio, R. Stevenson, B. Di Eugenio, and J. M. Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

S.P. Ponzetto and M. Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of HLT-NAACL*, pages 192–199.

B. Pouliquen and R. Steinberger. 2009. Automatic construction of multilingual name dictionaries. In *Learning Machine Translation*. MIT Press, NIPS series.

M. Recasens and E. Hovy. 2010. Coreference resolution across corpora: Languages, Coding schemes, and Preprocessing Information. In *Proceedings of ACL*, pages 1423–1432.

M. Recasens, L. Marquez, E. Sapena, M.A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley. 2010. SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. In *Proceedings of ACL*, pages 1–8.

J. Steinberger, M. Poesio, M. Kabadjov, and K. Ježek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680.

V. Stoyanov, N. Gilbert, C. Cardie, and E. Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of ACL-IJCNLP*.

M. Turchi, J. Steinberger, M. Kabadjov, and R. Steinberger. 2010. Using parallel corpora for multilingual (multi-document) summarisation evaluation. In *Proceedings of CLEF*, pages 52–63.

Y. Versley, S.P. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang, and A. Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Proceedings of LREC*.