Modelling Entity Instantiations

Andrew McKinlay
School of Computing
University of Leeds, UK
scs4ajm@comp.leeds.ac.uk

Katja Markert
School of Computing
University of Leeds, UK
markert@comp.leeds.ac.uk

Abstract

We introduce the problem of detecting *Entity Instantiations*, a type of entity relation in which a set of entities is introduced, and either a member or subset of this set is mentioned afterwards. We perform the first, reliable, corpus study of Entity Instantiations, concentrating on intersentential annotation. We then develop the first automatic instantiation detector, which incorporates lexical, contextual and world knowledge and shows significant improvements over a strong baseline.

1 Introduction

In this paper we annotate and classify *Entity Instantiations*. An Entity Instantiation is a non-coreferent entity relationship, where a *set* of entities is mentioned, and then a *member* or *subset*[1] of this set is introduced. Example 1 shows a pair of sentences with the set in bold and set member in italics.[2] Examples 2 and 3 show a pair of sentences with a set in bold and subset in italics.

(1) a. **Some European funds** recently have skyrocketed.
b. *Spain Fund* has surged to a startling 120% premium.

(2) a. **Bids totalling $515 million** were submitted.
b. *Accepted offers* ranged from 8.38% to 8.395%

(3) a. In the aftermath of the downturn **many manufacturers** have struggled.
b. *Those relying on foreign imports* have had the most difficulty.

The detection of Entity Instantiations is not tackled in ACE (ACE, 2000–2005) or MUC (MUC, 1987–1998), the two most popular schemes of semantic relation annotation. It is, however, important as it can supplement knowledge about the member or subset. In Example 4 below, the Entity Instantiation between *'several EU countries'* and *'the UK'* gives us the knowledge that not only are interest rates dropping in the UK, but inflation is rising as well. Entity Instantiations can also aid the interpretation of sentiment — in Example 5, the author's thoughts about the pay of Wayne Rooney can be inferred from the negative sentiment of the first sentence. In some instances, the member or subset is even uninterpretable without the set. In Example 3, *'Those relying on foreign imports'* requires *'many manufacturers'* to interpret the missing head noun. The problem of detecting these types of Entity Instantiation overlaps with bridging anaphora.

(4) a. Inflation has increased sharply in **several EU countries**.
b. In *the UK*, this has accompanied a drop in interest rates.

(5) a. **Footballers** are vastly overpaid.
b. Manchester United pay *Wayne Rooney* £200,000 per week.

The interpretation of Entity Instantiations can often be difficult. Entity Instantiations occur in a variety of forms. Participating noun phrases (NPs) include pronouns and proper nouns, can

[1] When we refer to a subset, we mean a *proper* subset. We consider two equal sets to be coreferent, and not participating in an Entity Instantiation.

[2] Examples 1, 2, 3, 8 and 9 are adapted from the Penn Treebank Wall Street Journal Corpus (Marcus et al., 1993).

Proceedings of Recent Advances in Natural Language Processing, pages 268–274,
Hissar, Bulgaria, 12-14 September 2011.

have missing head nouns (see Example 3) and fulfil various grammatical roles in a sentence. The two participants in an Entity Instantiation can have word overlap (see Example 1) or synonymous head nouns (see Example 2), but are often not related in such a simple manner. For instance, in Example 5, one needs to know that Wayne Rooney is a footballer to identify the Entity Instantiation. Additionally, correct interpretation of an Entity Instantiation often needs contextual knowledge. In Examples 6 and 7, the contextual information about the attitudes of the workers is necessary to establish whether an Entity Instantiation exists.

(6) a. **Some workers** are opposed to strike action.

b. *John Smith* fears that a strike could damage the industry's public perception.

(7) a. **Some workers** are opposed to strike action.

b. *David Jones*, however, is willing to put his job on the line for the cause. *(Not an instantiation.)*

In this paper we present an annotated corpus of Entity Instantiations, containing 648 annotated instantiations over 25 texts. We then use this corpus to train and test an automatic Entity Instantiation identifier, which gains significant improvements over a unigram baseline.

2 Related Work

Our work is related to Relation Extraction (RE), which is the discovery of semantic relations between pairs of entities. Much of the work in this field is connected to the Message Understanding Conferences (MUC, 1987–1998) and the NIST Automatic Content Extraction (ACE, 2000–2005) programs, both of which provide annotated corpora of semantic relations. The ACE-2004 scheme includes 7 broad relation types, divided into a total of 23 subtypes, such as *ART.User-Owner* to indicate the ownership of an object by a person, and *ORG-AFF.Employment* to represent the employment of a person by an organisation.

Entity Instantiations are not considered in the MUC and ACE annotation schemes, which consider relationships between different *types* of entity, such as those between persons and locations, rather than our groups and instances of entities of the same type. However, the algorithms used to classify these semantic relationship might still be applicable to our problem.

A variety of automatic RE algorithms have been developed, falling largely into two groups; those that learn from tree-kernels and those that use traditional, flat features. In one approach of the first type, (Zhou et al., 2007) use tree kernels to capture the structured information held in the parse trees of entities. They implement an algorithm which dynamically decides how much context to include as part of the tree, and in conjunction with some flat features it achieves an F-score of 75.8% on the 7 broad relation types in the ACE-2004 dataset.

Two recent flat-featured approaches successfully exploit background knowledge to improve RE. (Chan and Roth, 2010) implement features which use Wikipedia queries to search for *parent-child* relationships between entities. They attain an F-score of 68.2% at the coarse-grained level and 54.4% at the fine-grained level on a set of directed, sentence-internal relations from the ACE-2004 dataset. (Sun et al., 2011) generate large-scale word clusters from the TDT5 corpus and incorporate information regarding which cluster the mention head word belongs to. This method results in an F-score of 71.5%.

Our work is also related to the problem of bridging anaphora. A bridging anaphor is an anaphor that is not coreferent to its antecedent, but connected by another relationship, such as meronymy. Prior work in theoretical linguistics and corpus linguistics (Asher and Lascarides, 1998; Fraurud, 1990; Poesio and Vieira, 1998) has offered significant insight into bridging. A number of bridging publications also refer to set membership or subset relationships specifically (Clark, 1975; Prince, 1981; Gardent et al., 2003). Further work has concentrated on the development of algorithms for the resolution of bridging anaphora. (Markert et al., 1996; Vieira and Poesio, 2000) create end-to-end systems for bridging resolution, while both (Markert et al., 2003) and (Poesio et al., 2004) tackle solely part-of bridging references.

Our work differs from bridging in that often Entity Instantiations are not anaphoric (see Examples 1, 4, 5 and 6). There is, however, some overlap. For instance, in Example 3 the subset *'Those relying on foreign imports'* requires knowledge of the set *'manufacturers'* to be understood.

Our work is also related to (Recasens et al., 2010), in which the authors develop a typology

of near-identity coreference relationships, including largely overlapping sets. Set membership relations, however, are not tackled.

3 Corpus Study

To create a gold standard corpus creation we annotate full texts from the Penn Treebank (PTB) Wall Street Journal corpus (Marcus et al., 1993) for the presence of two types of Entity Instantiation:

Set Member A set of entities is introduced, and a *single member* of that set is mentioned.

Subset A set of entities is introduced, and a smaller *subset* of these is mentioned.

We limit our annotation to instantiations that occur *between* adjacent sentences. We do not annotate intrasententially, as we suspect that many intrasentential instantiations may be easily discoverable by syntactic analysis (for example, the instantiations in *'Some football managers, such as Sir Alex Ferguson'* and *'Among these workers, John Smith'*)..

Our annotation tool automatically identifies plural and singular noun phrases (NPs) that are candidates for participating in Entity Instantiations, separately displaying plural-plural NP pairs for subset annotation and plural-singular NP pairs for set member annotation. We automatically remove NPs that are appositions or predicates, and therefore not mentions. Our tool also includes the option to manually mark noun phrases as *"Not a mention"*. We use this to exclude instances of non-referential *it*, noun phrases that are idiomatic — such as *pie in the sky* — and generic pronouns.

The annotator then indicates whether each pair of NPs forms an Entity Instantiation. We annotate each pair of sentences twice; once with potential sets in first sentence and potential set members and subsets in the second sentence, and once with potential sets in the second sentence and potential set members and subsets in the first sentence.

3.1 Agreement Study

To ascertain the reliability and replicability of our annotations, we undertook a short agreement study. Five texts containing a total of 6,177 NP pairs were independently annotated by the two authors of this study, and their agreement was measured in the following three variations:

1. Does this pair of candidate noun phrases participate in a set membership/subset relationship or not?
2. Does this candidate set member/subset participate in a set membership/subset relationship with any potential set or not?
3. Is there an Entity Instantiation between these two sentences?

Method	# of items tested	Kappa	Agreement
1	6177 pairs of NPs	0.6504	97.31%
2	2994 NPs	0.6403	95.23%
3	607 sentence pairs	0.7317	91.09%

Table 1: Agreement Statistics

The results of the agreement study, including percentage agreement and chance corrected agreement (Kappa, (Cohen, 1960)), are presented in Table 1. Our agreement about which candidates were *"Not a mention"* was $\kappa = 0.7146$. These agreement statistics show reasonable agreement on the task, and that our annotation scheme is reliable and replicable.

There were several re-occurring types of disagreements. It was often difficult for annotators to establish whether a pair of sets were subsets, coreferent or overlapping. In Example 8, one can interpret *'men'* to mean either the men belonging to Baker or the general set of men, and this interpretation directly affects whether *'them'* is considered a subset.

Another problematic issue was systematic polysemy. In Example 9, *'Most cosmetic purchases'* might comprise a set of transactions or a set of products. The result of this interpretation then affects whether one considers *'lipstick'* to be a set member.

We also found that disagreements often propagated. A single decision about the relationship between two entities early on in a text can result in a large number of follow-on disagreements.

(8) a. Baker had lots of **men**.
    b. But she didn't trust *them* and didn't reward trust.

(9) a. **Most cosmetic purchases** are unplanned.
    b. *Lipstick* is often bought on a whim.

3.2 Further Annotation

After the successful agreement study, a further 20 texts were annotated by the first author of this study in order to complete the corpus. The frequency of Entity Instantiations over the final 25

Entity Instantiation	# NP pairs	%
Set Member	468	1.616
Subset	180	0.621
No inst. plural-singular	18758	64.76
No inst. plural-plural	9560	33.00
Total	28966	100

Table 2: Frequency of Entity Instantiations in 25 texts

texts is shown in Table 2. We found that a mean of 26 instantiations occurred per text, and that set membership instantiations occur considerably more frequently than subset instantiations.

4 Automatic Instantiation Detection

We use a supervised machine learning approach to detect which NP pairs comprise Entity Instantiations. Below we detail our feature set, experimental set-up and results.

4.1 Features

Our features fall into five broad categories; *surface*, *salience*, *syntactic*, *contextual* and *knowledge*. These categories contain both features that pertain to a single NP, and those that represent cross-NP relationships.

Surface features. Our surface features consist of unigrams, part-of-speech tags, lemmas, and dependency-parse[3] derived heads of each NP. We calculate Levenshtein's distance between the strings representing the unigrams, lemmas, head word and head lemma of each NP, hoping to capture pairs like *'funds'* and *'fund'* (see Example 1). We also calculate the distance in characters and words between NP pairs, and include these along with versions normalised by the total length of the two sentences containing the NPs. Additionally we include a boolean feature which represents the order of the NPs — True for candidate set NP in the first sentence and candidate set member/subset NP in the second sentence and False for the reverse order.

Salience features. As an indicator of the salience of each NP we include: its grammatical role, derived from dependency parse data; whether it is the first mention of that entity in the sentence or document; the number of mentions of the entity prior to this in the document; and the overall number of mentions of the entity in the document. We approximate the number of entity mentions by judging noun phrases with identical heads to be coreferent, as in (Barzilay and Lapata, 2008).

Syntactic features. We include five syntactic features, representing syntactic parallelism and pre- and post-modification. The modification type includes values that represent apposition, conjunction, pre modification and bare nouns. Our intuition is that set members and subsets are often more heavily modified than the sets that they are part of, as in *footballers → footballers playing in the Premiership*, *European countries → European nations that use the Euro*.

Contextual features. We include several contextual features, hypothesising that NPs that occur in similar contexts may be more likely to be Entity Instantiations. We retrieve the Levin class (Levin, 1993) of each NP's head verb, as well as the verb itself, noting examples such as Example 1 which has two similar verbs, *'surge'* and *'skyrocket'*. We also calculate whether each NP is in a quotation, and include an approximation of the discourse relations present in the two sentences by identifying likely discourse connectives and mapping them to their most frequent explicit relation in the Penn Discourse Treebank (PDTB) (Prasad et al., 2008). In cases such as Example 7, the presence of the discourse connective *'however'* appears useful in establishing that no instantiation is present. Note that we do *not* use any PDTB annotations to discover the presence of implicit or explicit discourse relations in the two sentences.

Knowledge-based features. Our knowledge-based features are organised into four categories:

WordNet. We use WordNet to establish whether the head words of NPs that are *not* named entities are synonyms or hyponyms, in an effort to identify pairs such as *'offers'* and *'bids'* in Example 2.

Freebase. We use Freebase (Bollacker et al., 2008), a freely-available repository of structured knowledge, to attempt to establish the relatedness of NPs. Each entity in Freebase is associated with a list of topics, which loosely represent hyponyms of the entity. For example, the topics listed for *'Wayne Rooney'* include *['Person','Football player','Athlete','2010 World Cup Athlete']*. For each NP representing a potential set

[3] Our dependency parses are generated from the gold standard PTB tree.

member or subset, we search Freebase using their Search API, choosing those matching entities that have a relevance score over 35. We then retrieve a list of topics for each entity and compare these topics to our potential set NP. If one of the topics is equal to, synonymous with, or has a Levenshtein distance of 1 from our potential set, the feature is True. Otherwise the feature is False.

Google PMI. We also use Google for discovering potential set membership and subset relations. We calculate Point-wise Mutual Information from hit counts for our potential Entity Instantiations, based on the notion that the pattern "X and other Y", where X is a potential set member or subset and Y is a potential set, indicates hyponymy (Hearst, 1992; Markert and Nissim, 2005). We use the following formula to calculate the value of our feature:

$$\text{G-PMI}(X, Y) = \frac{hits(\text{“}X \text{ and other } Y\text{”})}{hits(\text{“}X\text{”}) \times hits(\text{“and other } Y\text{”})}$$

Animacy. We attempt to establish whether the animacy of the two NPs match, reasoning that pairs of NPs that do not have the same animacy are highly unlikely to participate in an Entity Instantiation.

We use a list of animate pronouns, lists of animate and inanimate words distributed as part of the Stanford Deterministic Coreference Resolution System (Ji and Lin, 2009; Lee et al., 2011), and named entity information generated by the Stanford Named Entity Recognizer (Finkel et al., 2005) to ascertain the animacy of each NP. Our feature has three possible values; Match if the two NPs have the same animacy, No Match if they do not, and Not Present if we cannot calculate the animacy of one of the NPs. Not Present occurs in only 6% of pairs.

4.2 Experimental Set-up and Results

We divide our data set into two; plural-plural NP pairs that are labelled either *subset* or *no-instantiation* and plural-singular NP pairs that are labelled either *set member* or *no-instantiation*. We use the machine learner ICSIBoost (Favre et al., 2007). ICSIBoost is an open source implementation of Boostexter (Schapire and Singer, 2000), an algorithm which combines simple 'rules-of-thumb' — in this case, decision stumps — to produce a classifier. We apply 10-fold cross-validation for testing and training in all our experiments, keeping pairs from the same text in the same fold, to avoid rewarding the learning of very specific rules about the unigrams present which will not generalise well.

Due to the nature of the annotation study, there are many more pairs of candidates between which no Entity Instantiation has been annotated than those that have. Only 2.32% of the 28,966 pairs of candidates in the corpus have a set member or subset annotation. We therefore experiment with two different datasets.

Firstly, we used random sub-sampling to produce a balanced data set in which only 50% of the annotated pairs were non-relations, and used this for both training and testing. Results on the sub-sampled data are shown in Table 3.

Secondly, we experimented with the original, highly skewed data. Training on the original data resulted in a classifier that almost never predicted an instantiation, so we experimented with some simple techniques to improve precision and recall. These comprised randomly subsampling the negative examples so that they made up 50% or 75% of the training data, and oversampling the positive examples in the training data by a factor of 10, 20 or 40. The results of these experiments are shown in Table 4.

For comparison, results for a baseline whose sole features are the unigrams of the two NPs are also included. The Precision, Recall and F-Measure scores shown are for the positive examples in each set.

4.3 Discussion

On a balanced data set, our best features show highly significant improvements over the unigram baseline[4]. We performed a feature ablation study, removing each group of features from our model in turn, the results of which are present in Table 3. Our knowledge-based features are particularly good for identifying instantiations. Upon further investigation, we discovered that our Google PMI feature is the most effective of this feature group, with large PMI values often being indicative of instantiations.

Our salience features aid classification significantly for set members but not subsets. This indicates that set members are often first mentions of an entity that are mediated from a set, but subsets function less often in this way. In general, sub-

[4] $p < 10^{-}8$ and $10^{-}4$ for set members and subsets respectively with McNemar's χ^2 test (McNemar, 1947).

	Set Members				Subsets			
Feature set	**Accuracy**	**P**	**R**	**F**	**Accuracy**	**P**	**R**	**F**
Majority	50.0%	—	—	—	50.0%	—	—	—
Unigrams	58.8%	0.692	0.316	0.434	52.9%	0.565	0.255	0.352
All	68.9%♣	0.782	0.525	0.628	65.2%	0.724	0.489	0.584
All - Surface	66.6%	0.717	0.550	0.622	62.00%	0.651	0.516	0.576
All - Salience	65.5%♠	0.739	0.479	0.582	65.4%♣	0.730	0.489	0.586
All - Syntax	68.0%	0.770	0.512	0.615	65.2%	0.732	0.479	0.579
All - Contextual	67.7%	0.792	0.479	0.597	63.0%	0.674	0.505	0.578
All - World Knowledge	64.4%◇	0.766	0.413	0.537	60.6%♠	0.675	0.410	0.510

Table 3: Results on balanced data set

♣ Algorithm with highest accuracy
♠ Significantly *worse* than ♣, significance $p < 0.005$, McNemar's χ^2 test.
◇ Significantly *worse* than ♣, significance $p < 0.001$, McNemar's χ^2 test.

	Set Members				Subsets			
Method	**Accuracy**	**P**	**R**	**F**	**Accuracy**	**P**	**R**	**F**
Original Set	97.39%	0.2979	0.0289	0.0527	97.90%	0.1852	0.0266	0.0465
Undersampling 50/50	83.31%	0.0782	0.5227	0.1361	76.47%	0.0453	0.5585	0.0839
Undersampling 75/25	94.60%	0.1275	0.1963	0.1546	93.28%	0.0838	0.2500	0.1255
Oversampling x10	96.89%	0.2500	0.1178	0.1601	97.47%	0.1685	0.0798	0.1083
Oversampling x20	96.38%	0.2129	0.1632	0.1848	97.21%	0.1557	0.1011	0.1226
Oversampling x40	95.24%	0.1690	0.2272	0.1938	96.51%	0.1346	0.1489	0.1414

Table 4: Results on unbalanced data set

sets appear harder to detect than set membership relations, but the smaller size of the subset data set likely contributes to this.

Learning from the original, highly skewed data is much more difficult, and our highest F-scores are 0.1938 and 0.1414 for set members and subsets, respectively (see Table 4). Learning from data with this sort of distribution is difficult, regardless of the domain. In future we intend to use techniques such as SMOTE (Chawla et al., 2002) and One-Sided Selection (Kubat and Matwin, 1997) to address this heavy skew.

5 Conclusion and Future Work

We propose a novel Information Extraction task: the detection of Entity Instantiations. This task is potentially important for a variety of NLP problems, such as question answering and sentiment analysis. We have presented the first corpus study of Entity Instantiations, achieving good levels of annotator agreement. Our supervised machine learning classifier achieves an F-score of 0.628 for set member relations and 0.586 for subset relations on a balanced set, making good use of a variety of features, including world-knowledge and salience criteria.

In the future, we intend to expand our annotation to include intrasentential and further distant Entity Instantiations, as well as our current instantiations between adjacent sentences. Future machine learning approaches to consider are tree-kernel based approaches such as (Zhou et al., 2007). To tackle the high skew in our data, we will use techniques such as those detailed in (Kubat and Matwin, 1997) and (Chawla et al., 2002), and also look to methods such as active learning to acquire more positive instantiation examples.

Acknowledgements

Andrew McKinlay is funded by an EPSRC Doctoral Training Grant. This research draws on data provided by the University Research Program for Google Search, a service provided by Google to promote a greater understanding of the web.

References

ACE. 2000-2005. Automatic Content Extraction. `http://www.ldc.upenn.edu/Projects/ACE/`.

N. Asher and A. Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1):83–113.

R. Barzilay and M. Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250.

Y.S. Chan and D. Roth. 2010. Exploiting background knowledge for relation extraction. In *Proceedings of COLING 2010*, pages 152–160.

N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. 2002. Smote: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357.

H.H. Clark. 1975. Bridging. In *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, pages 169–174.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

B. Favre, D. Hakkani-Tür, and S. Cuendet. 2007. ICSIboost. `http://code.google.come/p/icsiboost`.

J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of ACL 2005*, pages 363–370.

K. Fraurud. 1990. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7(4):395.

C. Gardent, H. Manuelian, and E. Kow. 2003. Which bridges for bridging definite descriptions. In *Proceedings of the EACL 2003 Workshop on Linguistically Interpreted Corpora*, pages 69–76.

M.A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING 1992*, pages 539–545.

H. Ji and D. Lin. 2009. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proceedings of PACLIC 2009*.

M. Kubat and S. Matwin. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of ICML 1997*, pages 179–186.

H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. 2011. Stanfords multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the CoNLL-2011 Shared Task*, pages 28–34.

B. Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

K. Markert and M. Nissim. 2005. Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, 31(3):367–402.

K. Markert, M. Strube, and U. Hahn. 1996. Inferential realization constraints on functional anaphora in the centering model. In *Proceedings of CogSci 1996*, pages 609–614.

K. Markert, N. Modjeska, and M. Nissim. 2003. Using the web for nominal anaphora resolution. In *Proceedings of EACL 2003 Workshop on the Computational Treatment of Anaphora*, pages 39–46.

Q. McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

MUC. 1987-1998. The NIST MUC website: `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/`.

M. Poesio and R. Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24:183–216, June.

M. Poesio, R. Mehta, A. Maroudas, and J. Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of ACL 2004*, page 143.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 2008*, pages 2961–2968.

E.F. Prince. 1981. Toward a Taxonomy of Given-New Information. *Radical Pragmatics*, 3:223–255.

M. Recasens, E. Hovy, and M.A. Martı. 2010. A typology of near-identity relations for coreference (NIDENT). In *Proceedings of LREC 2010*, pages 149–156.

R.E. Schapire and Y. Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine learning*, 39(2):135–168.

A. Sun, R. Grishman, and S. Sekine. 2011. Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of ACL-HLT 2011*, pages 521–529.

R. Vieira and M. Poesio. 2000. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

G.D. Zhou, M. Zhang, D.H. Ji, and Q.M. Zhu. 2007. Tree Kernel-Based Relation Extraction with Context-Sensitive Structured Parse Tree Information. In *Proceedings of EMNLP-CoNLL 2007*, pages 728–736.