Investigating Features for Classifying Noun Relations

Dominick Ng, David J. Kedziora, Terry T. W. Miu and **James R. Curran**
School of Information Technologies
University of Sydney
NSW 2006 Australia
{dong7223,dked9370,tmiu9098,james}@it.usyd.edu.au

Abstract

Automated recognition of the semantic relationship between two nouns in a sentence is useful for a wide variety of tasks in NLP. Previous approaches have used kernel methods with semantic and lexical evidence for classification. We present a system based on a maximum entropy classifier which also considers both the grammatical dependencies in a sentence and significance information based on the Google Web 1T dataset.

We report results comparable with state of the art performance using limited data based on the SemEval 2007 shared task on nominal classification.

1 Introduction

Analysis of the semantics of natural language is an area of research undergoing renewed interest, driven by the many applications which can directly benefit from such information, including question answering and text summarization. In particular, recognition of the relationship between words in a sentence is useful for clarifying ambiguities in tools that attempt to interpret and respond to natural language. Recent developments in the semantic web – a vision where information is comprehensible to machines as well as people – have also accelerated the need for tools which can automatically analyse nominal relationships.

We approach the task of relation classification using the relation set and training data provided by the SemEval 2007 shared task on nominal classification (Girju et al., 2007). The problem as defined by the task description is to discover the underlying relationship between the concepts expressed by two nominals, excluding named entities. The relationship is informed by the context of an English sentence, e.g. it is clear that the relationship between door and car differs in the fragments the car door and the car scraped the door of the garage. Resolving the relation between nominals in cases where ambiguities exist is useful for generalisation in NLP systems.

We created a system based upon a maximum entropy (ME) classifier developed by Clark and Curran (2004). A separate binary classifier for each of the relations was trained over the corresponding training data, and the additional features used for each relation were selected by performing a seven-fold cross-validation over all combinations of features developed for this task. We report an overall accuracy of 71.9% macro-averaged over the seven relations and an overall F-measure of 70.7%, comparable with state of the art performance.

2 Background

The task of relation classification is complicated by the lack of consensus on relation sets and algorithms. Previous research has studied areas as diverse as noun compound classification in the medical domain (Rosario and Hearst, 2001), gene relations (Stephens et al., 2001), verb-verb semantic relations (Chklovski and Pantel, 2004), and noun-modifier relations (Nastase and Szpakowicz, 2003). Many independent class hierarchies have been developed to suit each application domain, and it is difficult to transfer one

Relation	Training Example
Cause-Effect	[e_1 Famine] following [e_2 drought] has hit the West African savannahs, where there have been other bad droughts.
Instrument-Agency	The [e_1 judge] hesitates, [e_2 gavel] poised.
Product-Producer	The [e_1 artist] made the [e_2 picture] when he was in fourth grade.
Origin-Entity	It's unfortunate you didn't try a [e_1 potato] [e_2 vodka].
Theme-Tool	The [e_1 submission] [e_2 deadline] is February, 2, 2007.
Part-Whole	Typically, an unglazed [e_1 clay] [e_2 pot] is submerged for 15 to 30 minutes to absorb water.
Content-Container	The [e_1 kitchen] holds patient [e_2 drinks] and snacks.

Table 1: Examples of the SemEval relations.

of these hierarchies to another domain. The organisers of the SemEval task defined the classification problem in terms of seven semantic relations commonly mentioned by researchers, and a list of these along with some training examples is provided in Table 1. An annotated dataset of 140 training examples and at least 70 test sentences was created for each relation by searching the web using wild-card search patterns satisfying the constraints of each relation, e.g. * holds * for the Content-Container relation. This method was used in order to provide near miss negative examples (Girju et al., 2007).

Fifteen systems split into four categories were submitted for the SemEval 2007 workshop. Almost all of the systems utilised extended feature sets that built upon the data provided by the task; most systems also implemented some form of statistical or kernel approach to develop binary classifiers for the relations (see Bedmar et al. (2007), Hendrickx et al. (2007), and Nulty (2007) for some previous approaches to the classification task explored in this paper). The best performing systems achieved F-measures in the range of 71.5% – 72.4% by utilising the provided WordNet sense keys and adding more training examples to those supplied by the task; however, the majority of systems did not augment the provided data like this and reported F-measures and accuracies below 67.2% (Girju et al., 2007).

Each training example in the annotated dataset consists of a sentence, two nominals whose relationship is to be evaluated, WordNet 3.0 sense keys for each of the nominals, the wild-card query used to obtain the example, and comments on the choices made during the creation of the example. The evaluation drew distinction between systems that did and did not use the supplied WordNet and wild-card query information.

3 Maximum Entropy Modelling

The entropy of a classifier is a measure of how predictable that classifier's decisions are. The lower the entropy, the more biased a classifier is, i.e. a relation classifier has zero entropy if it always assigns the same relation to any input. The theory underpinning ME modelling is that the distribution chosen to fit the specified constraints will eliminate biases by being as uniform as possible. Such models are useful in NLP applications because they can effectively incorporate diverse and overlapping features whilst also addressing statistical dependencies.

We used the ME implementation described in Clark and Curran (2004). The ME models used have the following form:

$$p(y|x,\lambda) = \frac{1}{Z(x|\lambda)} \exp\left(\sum_{k=1}^{n} \lambda_k f_k(x,y)\right)$$

where $Z(x|\lambda)$ is the normalisation function and the f_k are features with associated weights λ_k. The system uses Gaussian smoothing on the parameters of the model. The features are binary-valued functions which pair a relation y with various observations x from the context provided, e.g.

$$f_j(x,y) = \begin{cases} 1 & \text{if } word(x) = damage \ \& \\ & y = \text{Cause-Effect-True} \\ 0 & \text{otherwise} \end{cases}$$

4 Features and Methodology

We focused our efforts on finding features which aggressively generalise the initial material over as broad a search space as possible. We investigated lexical, semantic, and statistical features sourced from a number of corpora as well as morpho-syntactic features from a grammatical parse of each sentence. Features were evaluated using a seven-fold cross-validation performed over the training data for each relation over every possible combination of features – a process made possible by the small size of the corpora and the relatively small number of features experimented with. The speed of the training process for the ME implementation was also a factor in enabling the exhaustive search.

The features which resulted in the best performance for each relation in the cross-validation were then used to train seven binary classifiers for the final run over the supplied test data.

4.1 Preprocessing

Prior to feature generation our system extracted from the supplied data the sentences and marked nominals (termed as e_1 and e_2). While WordNet was used internally as features by our system, we did not use the specific sense keys provided by the data to query WordNet – we relied upon a more general word lookup that extracted all of the possible senses for the nominals. We also did not make use of the provided query, based on its ineffectiveness in previous studies (Girju et al., 2007).

Close examination of the training data also revealed some negative training examples that were identified in comments as belonging to a different relation set. These examples were collected and added to the appropriate training file to further extend the original dataset. However, no new examples were created: only examples which had already been identified as belonging to a particular relation were added in this process.

4.2 Lexical Features

Lexical features are useful for capturing contextual information about the training example, and they are the most obvious features to incorporate. However, due to the limited amount of training data available for this task, lexical features encounter sparseness problems as there are few relevant collisions between words. We utilised the following lexical features:

- **sen**: The words of the sentence itself. This was used as the baseline for feature testing;
- **red**: A reduced version of the sentence with all words of length 2 or less removed;
- **heads**: The head words of the nominals in question, e.g. for [e_1 tumor shrinkage] after [e_2 radiation therapy] the relation actually holds between shrinkage and therapy. This feature is very specific, but allows for nominals which are commonly linked to certain relations to be identified;
- **dir**: The required direction of the relation (i.e. from e_1 to e_2 or vice versa) that is encoded in the data – useful as some relations are more likely to exist in a particular direction, e.g. the Part-Whole relation is most commonly found encoded in the direction [e_1 Part]-[e_2 Whole] (Beamer et al., 2007).

4.3 WordNet Features

WordNet (Fellbaum, 1998) is the most heavily used database of lexical semantics in NLP. Created at Princeton University, WordNet is based around groups of synonyms (synsets) and encodes a vast array of semantic properties and relationships between these synsets. The coverage of WordNet means that it is very useful for generalising features over a small corpus of data, and many previous approaches to classification tasks have utilised WordNet in some way – including most of the systems from the SemEval proceedings (Girju et al., 2007). However, unlike most of these systems, we did not use the supplied WordNet sense keys as we believe that it is unrealistic to have such precise data in real-world applications. As a consequence, all of our WordNet features were extracted using the indicated nominals as query points.

- **syn**: Synonyms of the nominals. We extracted from WordNet all synonyms in all senses for each of the marked nouns;
- **hyp1, hyp2**: Hypernyms of the nominals, i.e. more general concepts which encompass the nominals. These features allow us to

broaden the coverage given by the nominals over less specific entities. We exhaustively mined all hypernyms of the marked nouns to a height of two levels, and encoded the two levels as separate features;

- **lex**: Lexical file numbers, which correspond to a number of abstract semantic classes in WordNet, including noun.artifact, noun.event, and noun.process. This allows for nominal relations which do not make sense to be identified, e.g. a noun.process should not be able to contain a noun.event, but the process may cause the event (Bedmar et al., 2007);

- **cont**: Container - a binary feature indicating whether the marked nouns are hyponyms (more specific concepts) of the container synset. This feature was included mainly for the benefit of the Content-Container relation; however, we hypothesised that their inclusion may also assist in classifying other relations; e.g. the 'effect' in Cause-Effect should not be a physical entity.

4.4 Grammatical Relations Features

Syntactic features representing the *path* between nominals are a useful complement for semantic and lexical features because they account for the way in which words are commonly used in text. Semantic relationships can often be associated with certain patterns of words, e.g. the pattern e_1 is inside e_2 is a strong indicator for the Content-Container relation for many general combinations of e_1 and e_2. However, these patterns can be expressed in many different ways - inside e_2 e_1 is or inside e_2 is e_1 are other ways of expressing a Content-Container relationship – and while the words are essentially the same between the examples the changed ordering creates difficulties in designing good features. This problem can be alleviated by considering syntactic dependencies in a sentence rather than a naive concatenation of words (Nicolae et al., 2007).

Grammatical relations (GRs) represent the syntactic dependencies that hold between a head and a dependent in text. Initially proposed by Carroll et al. (1998) as a framework-independent metric for parsing accuracy, GRs are arranged in a hierarchy that allows for varying levels of exactness in parsing: a general *dependent* relation can be assigned to indicate that there is some doubt over the precise dependency that holds between two words. We postulated that a simple graph constructed from the dependencies (whereby words of the text are nodes and undirected edges are added between nodes if there is some grammatical relation that links them) could be used to find a path between the two nominals in each sentence. This path would compare favourably to a naive concatenation of the words between the nominals as it considers the actual dependencies in the sentence rather than just the positions of the words, although in many cases at least one of the words between the marked nominals in the sentence will be represented in the dependency path. Table 2 gives a list of GRs used in this process.

GRs	Description
conj	coordinator
aux	auxiliary
det	determiner
ncmod	non-clausal modifier
xmod	unsaturated predicative modifier
cmod	saturated clausal modifier
pmod	PP modifier with a PP complement
ncsubj	non-clausal subject
xsubj	unsaturated predicative subject
csubj	saturated clausal subject
dobj	direct object
obj2	second object
iobj	indirect object
pcomp	PP which is a PP complement
xcomp	unsaturated VP complement
ccomp	saturated clausal complement
ta	textual adjunct delimited by punctuation

Table 2: A list of GRs

```
(det man_1 A_0)
(ncmod _ does_2 not_3)
(aux talk_4 does_2)
(ncsubj talk_4 man_1 _)
(det woman_7 every_6)
(ncsubj walks_8 woman_7 _)
(conj or_5 walks_8)
(conj or_5 does_2)
```

Figure 1: GRs output from the C&C parser

To extract the grammatical relations from the provided data we parsed each training and test example with the C&C parser developed by Clark and Curran (2007). Figure 1 gives an example of the GRs for the sentence A man does not talk or every woman walks. A dependency graph was generated from this output and the shortest path between the nominals found. In the example in Figure 1, the path between man and woman is talk_does_or_walks.

Features were extracted from this path output in two formats: a generalised version (labelled with a 'g' prefix), whereby the two nominals in question were replaced whenever they appeared with the marker tags e_1 and e_2, and the actual version, where this extra generalisation step was not applied. We reasoned that the generalised output would be more useful as a classification feature as it removed the stipulation on the start and end of the path; however, we also felt that keeping the identity of the nominals would aid in classifying words often paired with prepositions that suggest some form of spatial or logical relationship, e.g. the fragment after_hurricanes suggests some form of Cause-Effect relationship from the temporal indicator 'after'. Our path features included:

- **path, gpath**: The path itself in a concatenated format, e.g. damage_comes_after_hurricanes or e_1_comes_after_e_2. These patterns were postulated to have some correlation with each relation;
- **strip, gstrip**: The path with a length filter of 2 applied;
- **slice, gslice**: The nominals with their immediate neighbour from the path, e.g. damage_comes, after_hurricanes or e_1_comes, after_e_2;
- **pair, gpair**: The bigrams in the path, e.g. e_1_comes, comes_after, after_e_2;
- **ptag, gptag**: The underscore-concatenated POS tags of the path words.

4.5 Web 1T Significance Features

Web 1T (Brants and Franz, 2006) is a Google-released corpus containing English word ngrams and their observed frequency counts in a body of approximately 1 trillion word tokens of text from publicly accessible web pages. Web 1T counts the occurrences of unigrams, 2-, 3-, 4-, and 5-grams in the 1 trillion word tokens, discarding unigrams appearing less than 200 times in the tokens (1 in 5 billion) and n-grams appearing less than 40 times (1 in 25 billion) in the tokens. This resource captures many lexical patterns used in common English, though there are some inconsistencies due to the permissive nature of the web: some commonly misspelt words are included and some text in languages other than English are also present.

O_{11}: freq count of word and bound together
O_{12}: freq count of bound without word
O_{21}: freq count of word without bound
O_{22}: freq count of neither bound or word
N: total number of tokens

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12}) \times (O_{11}+O_{21}) \times (O_{12}+O_{22}) \times (O_{21}+O_{22})}$$

Figure 2: The notation and formula used for the significance testing.

The idea of searching a large corpus for specific lexical patterns to indicate semantic relations of interest was first described by Hearst (1992). As previously mentioned, we postulated that certain patterns of words would associate with certain relations, but a naive concatenation of words located between the nominals would be unhelpful with such a small data set. This problem can be avoided by examining the frequencies of lexical patterns within a much larger dataset such as Web 1T, where the problem of data sparseness is offset by the size of the corpus. This pattern information would complement the semantic and syntactic information already used by incorporating evidence regarding word use in real-world text.

We chose to conduct statistical significance tests with the intention of observing if the presence of particular words between the nominals is meaningful, irrespective of whether or not they are in the sentences themselves. This allows us to collate all the words found to be significant when placed between the nominals and use them as ngram features. Our methodology is justified by the observation that patterns correlated with relations are likely to contain the same words regardless of the bounds, i.e. the pattern e_1 is inside

Baseline	F± std
sen	61.3±9.29
Feature	Mean Improv± std
red	+0.3±1.67
heads	+0.7±2.47
dir	+0.3±2.10
syn	+1.2±3.29
hyp1	+1.3±4.25
hyp2	+5.6±4.39
lex	+6.0±6.75
cont	+1.5±2.52
path	+0.2±1.00
gpath	+0.5±3.17
strip	+0.2±1.00
gstrip	-0.5±3.47
slice	+1.5±3.03
gslice	+0.5±2.88
pair	+0.2±1.93
gpair	+1.4±2.82
ptag	+0.5±1.54
gptag	-0.6±2.79
ngram	+1.8±2.94

Table 3: The average improvement in F-measure (using the words of the sentence as a baseline) for each feature macro-averaged over all 7 relations

e_2 is a strong indicator for the Content-Container relation for general combinations of e_1 and e_2.

The significance test was conducted as in Manning and Schütze (2000). The Web 1T data was searched for any 3-, 4-, and 5-grams that had the same bounds as the nominals of the sentence in question, i.e. patterns which match e_1 ... e_2. Then, for every intermediate word in the pattern, a χ-squared value was calculated to measure the significance of the word in relation to the bounds. This process was repeated for each training example, and Figure 2 gives the equations used for this test. We conducted some brief experiments to find the range of χ-squared values returned by this test; based on these we chose the χ-squared value of 10 to indicate significance to the training example being analysed, and selected all words with a χ-squared value above this level to add as ngram features.

5 Preliminary Feature Testing

As an initial step, we used the sentence words of each example as a baseline to test the individual performance of each feature in a seven-fold cross-validation over the training examples for each relation. We did this to compare the discrete improvement over the baseline that each feature offered and to allow a comparison as to how combining the features improves performance.

Relation	Best	Improvement
Cause-Effect	lex	+8.97
Instrument-Agency	hyp2	+7.31
Product-Producer	hyp1	+5.68
Origin-Entity	lex	+12.96
Theme-Tool	lex	+16.45
Part-Whole	hyp1	+2.95
Content-Container	hyp2	+6.84

Table 4: The best performing single features (using the words of the sentence as a baseline) and their mean improvement in F-measure for each relation

Table 3 shows that most features offer small gains on average over the baseline sentence, but also exhibit varying degrees of performance over the relations as seen in the relatively large standard deviations. In particular, lexical file numbers and second-level hypernyms have the largest mean improvement in F-measure, but also the largest standard deviations – indicating a widespread distribution of positive and negative contributions. Table 4 shows that these two features improve the baseline F-measure of five of the relations, implying from the large standard deviations that they severely worsen the performance of the remaining two. This behaviour is explained by noting the wide generalisation that these features add, creating the most collisions between training and test data and hence affecting the decisions of the classifier the most.

6 Results and Discussion

The features chosen to train the classifiers for the final system along with performance in the cross-validation are given in Table 5. All the relations performed best with a combination of lexical, semantic, and syntactic features, and three relations also used the statistical significance data obtained from Web 1T. The relatively even spread of feature types across relations implies that the classifier performs best when presented with a wide range of evidence that it can then combine into a model. However, the largest number of fea-

Relation	Features selected	F± std	Acc± std
Cause-Effect	dir, heads, cont, lex, pair, g/slice, g/strip, ngram	77.9± 7.06	73.6±9.00
Instrument-Agency	heads, cont, hyp2, lex, g/pair, gptag, gpath, g/slice, gstrip	78.2± 7.59	77.1±9.94
Product-Producer	heads, red, cont, hyp1, hyp2, gpath, pair, slice, strip	80.6± 3.67	72.1±3.93
Origin-Entity	dir, sen, hyp2, lex, gstrip	62.3±12.92	70.7±8.38
Theme-Tool	heads, lex, g/pair, gslice	71.2± 8.43	77.9±4.88
Part-Whole	dir, red, hyp1, syn, g/pair, slice, ngram	81.6± 8.81	77.1±8.09
Content-Container	heads, red, cont, hyp2, lex, pair, slice, ngram	71.1±12.22	72.9±6.36
Average	–	74.7± 6.88	74.5±2.85

Table 5: The best performing features with F-measure and accuracy percentages from the cross-validation

tures used was 11 – considerably less than the 20 tested during cross-validation – and this supports the general conclusion that using too many features in a maximum entropy approach with a small amount of training data adversely affects classifying performance.

The most commonly selected features were the Grammatical Relations bigram features (pair and g / slice). These features were used in all but one of the classifiers, indicating that bigram information provided very useful evidence for relation classification. Given that most path bigrams involve the nominals with a preposition that indicates a temporal or spatial relationship, we infer that the syntactic dependency between nominals and prepositions is an important feature for semantic relation classification. Other commonly selected features were the head words of the sentence and their lexical file numbers – these were present together in the Cause-Effect, Instrument-Agency, Theme-Tool, and Content-Container classifiers. This correlation is expected given that these relations usually exist between nominals that generally correspond with the semantic classes from WordNet.

Table 5 shows that some relations were more challenging to classify than others. Origin-Entity in particular exhibited the worst performance, with a standard deviation of 12.92 around an average F-measure of 62.3% under seven-fold cross-validation. This poor performance was expected given that most attempts from the SemEval proceedings rated Origin-Entity as equal hardest to classify along with Theme-Tool (Girju et al., 2007). On the other hand, our cross-validation yielded good performance for Theme-Tool, with an F-measure of 71.2% – potentially showing that maximum entropy methods are more effective at handling difficult relations than kernel approaches to the problem. Also notable are the strong performances of the Part-Whole and Product-Producer classifiers, with F-measures above 80% and accuracies above 72%. The other relations also performed well, with no other classifier exhibiting an F-measure or accuracy score below 70%.

Relation	P	R	F	Acc
Cause-Effect	78.0	69.6	73.6	71.3
Instrument-Agency	84.2	72.7	78.1	76.9
Product-Producer	87.1	70.1	77.7	66.7
Origin-Entity	50.0	75.0	60.0	70.4
Theme-Tool	62.1	72.0	66.7	74.6
Part-Whole	80.8	51.2	62.7	65.3
Content-Container	65.8	89.3	75.8	78.4
Average	**72.6**	**71.4**	**70.7**	**71.9**

Table 6: Final percentage precision, recall, F-measure, and accuracy results over the test data using the features listed in Table 5

Table 6 gives the final classifying results over the supplied test data using all the training examples and features selected in the cross-validation step as training material. We established a new benchmark for classifying the Instrument-Agency relation: our F-measure of 78.1% exceeds the best result of 77.9% for the relation from the SemEval proceedings (Girju et al., 2007). However, as a general rule, system performance was weaker over the test data than during the cross-validation step, providing some evidence of overfitting to the training data. This was particularly demonstrated in the markedly poor performance of the Part-Whole classifier – from the cross-validation F-measure dropped by 18.9% to 62.7% and accuracy fell 12.4% from 77.1% to 65.3%. It should

be noted however that our system performed better than most others in classifying the difficult relations as ranked in the SemEval task (Girju et al., 2007).

We recorded a final F-measure of 70.7% and accuracy of 71.9% macroaveraged over the seven relations, an improvement of 5.9% in both F-measure and accuracy over the best performing system using the same data (no WordNet sense keys or query) from SemEval 2007. Our system performed within an F-measure of 1.7% and accuracy of 4.4% of the top system from SemEval 2007, which incorporated a large number of extra training examples and WordNet sense keys (Beamer et al., 2007). Our results are comparable with more recent approaches to the same classification task, utilising pattern clusters (F-measure 70.6%, accuracy 70.1%, in Davidov and Rappoport (2008)) and distributional kernels (F-measure 68.8%, accuracy 71.4%, in Séaghdha and Copestake (2008)).

Overall these results show that a maximum entropy approach with a range of informative features is a feasible and effective method of classifying nominal relations when presented with limited data.

7 Conclusion

We have created a system built around a maximum entropy classifier that achieves results comparable with state-of-the-art with limited training data. We have also demonstrated that syntactic dependencies and frequency-based statistical features taken from large corpora provide useful evidence for classification, especially when combined with lexical and semantic information.

We have also shown that a maximum entropy approach using informative features performs strongly in the task of relation classification, and that exact WordNet sense keys are not necessary for good performance. This is important since it is impractical in large scale classifying tasks to provide this annotation.

The corpora is extremely small, and it should be noted that the choice to select the dataset using a limited number of queries artificially limits the scope of this task. We feel that an effort to annotate a large amount of randomly selected text with several hundred positive examples would greatly benefit further research into relation classification and validate the results presented in this paper.

Future improvements to the system could include incorporating more external resources (e.g. VerbNet), introducing Word-Sense Disambiguation as a replacement for WordNet sense keys, or by incorporating more relation-specific features, such as meronym (Has-Part) information from WordNet for the Part-Whole and Content-Container relations. More sophisticated analysis of the Web 1T data could also be undertaken, such as a generalised attempt to identify patterns underpinning semantic relationships, rather than just those corresponding to the provided sentences.

We achieved a final overall F-measure of 70.7% and accuracy of 71.9%, establishing a new benchmark for performance over the SemEval data without sense keys. Our system is also competitive with approaches that use sense keys, and so we expect that it will provide useful semantic information for classification and retrieval problems in the future.

Acknowledgments

The authors would like to thank the three anonymous reviewers, whose comments greatly improved the quality of this paper. Dominick Ng was supported by a University of Sydney Merit Scholarship; Terry Miu was supported by a University of Sydney Outstanding Achievement Scholarship and a University of Sydney International Merit Scholarship.

References

Brandon Beamer, Suma Bhat, Brant Chee, Andrew Fister, Alla Rozovskaya, and Roxana Girju. 2007. UIUC: A knowledge-rich approach to identifying semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 386–389, Prague, Czech Republic.

Isabel Segura Bedmar, Doaa Samy, and Jose L. Martinez. 2007. UC3M: Classification of semantic relations between nominals using sequential minimal optimization. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 382–385, Prague, Czech Republic.

Thorsten Brants and Alex Franz. 2006. Web 1T

5-gram version 1. Linguistic Data Consortium, Philadelphia. LDC2006T13.

John Carroll, Ted Briscoe, and Antonio Sanfilippo. 1998. Parser evaluation: a survey and a new proposal. In *Proceedings, First International Conference on Language Resources and Evaluation*, pages 447–454, Granada, Spain.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods on Natural Language Processing (EMNLP-2004)*, pages 33–40, Barcelona, Spain.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and log-linear models. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 104–111, Barcelona, Spain.

Stephen Clark and James R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Dmitry Davidov and Ari Rappoport. 2008. Classification of semantic relationships between nominals using pattern clusters. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 227–235, Ohio, USA.

Christiane Fellbaum, editor. 1998. *WordNet - An Electronic Lexical Database*. MIT Press.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-1992)*, pages 539–545, Nantes, France.

Iris Hendrickx, Roser Morante, Caroline Sporleder, and Antal van den Bosch. 2007. ILK: Machine learning of semantic relations with shallow features and almost no data. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 187–190, Prague, Czech Republic.

Chris Manning and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing*. MIT Press, Massachusetts, United States.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Proceedings of the 5th International Workshop on Computational Semantics*, pages 285–301, Tilburg, The Netherlands.

Cristina Nicolae, Gabriel Nicolae, and Sanda Harabagiu. 2007. UTD-HLT-CG: Semantic architecture for metonymy resolution and classification of nominal relations. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 454–459, Prague, Czech Republic.

Paul Nulty. 2007. UCD-PN: Classification of semantic relations between nominals using wordnet and web counts. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 374–377, Prague, Czech Republic.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-2001)*, pages 82–90, Pennsylvania, United States.

Diarmuid Ó Séaghdha and Ann Copestake. 2008. Semantic classification with distributional kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, pages 649–655, Manchester, UK.

Matthew Stephens, Mathew J. Palakal, Snehasis Mukhopadhyay, Rajeev R. Raje, and Javed Mostafa. 2001. Detecting gene relations from MEDLINE abstracts. In *Proceedings of the Sixth Annual Pacific Symposium on Biocomputing*, pages 483–496, Hawaii, United States.