SemEval-2007 Task 04: Classification of Semantic Relations between Nominals

Roxana Girju
Univ. of Illinois
at Urbana-Champaign
Urbana, IL 61801
girju@uiuc.edu

Preslav Nakov
Univ. of California at Berkeley
Berkeley, CA 94720
nakov@cs.berkeley.edu

Vivi Nastase
EML Research gGmbH
Heidelberg, Germany 69118
nastase@eml-research.de

Stan Szpakowicz
University of Ottawa
Ottawa, ON K1N 6N5
szpak@site.uottawa.ca

Peter Turney
National Research Council of Canada
Ottawa, ON K1A 0R6
peter.turney@nrc-cnrc.gc.ca

Deniz Yuret
Koç University
Istanbul, Turkey 34450
dyuret@ku.edu.tr

Abstract

The NLP community has shown a renewed interest in deeper semantic analyses, among them automatic recognition of relations between pairs of words in a text. We present an evaluation task designed to provide a framework for comparing different approaches to classifying semantic relations between nominals in a sentence. This is part of SemEval, the 4^{th} edition of the semantic evaluation event previously known as SensEval. We define the task, describe the training/test data and their creation, list the participating systems and discuss their results. There were 14 teams who submitted 15 systems.

1 Task Description and Related Work

The theme of Task 4 is the classification of semantic relations between simple nominals (nouns or base noun phrases) other than named entities – *honey bee*, for example, shows an instance of the Product-Producer relation. The classification occurs in the context of a sentence in a written English text. Algorithms for classifying semantic relations can be applied in information retrieval, information extraction, text summarization, question answering and so on. The recognition of textual entailment (Tatu and Moldovan, 2005) is an example of successful use of this type of deeper analysis in high-end NLP applications.

The literature shows a wide variety of methods of nominal relation classification. They depend as much on the training data as on the domain of application and the available resources. Rosario and Hearst (2001) classify noun compounds from the domain of medicine, using 13 classes that describe the semantic relation between the head noun and the modifier in a given noun compound. Rosario et al. (2002) classify noun compounds using the MeSH hierarchy and a multi-level hierarchy of semantic relations, with 15 classes at the top level. Nastase and Szpakowicz (2003) present a two-level hierarchy for classifying noun-modifier relations in base noun phrases from general text, with 5 classes at the top and 30 classes at the bottom; other researchers (Turney and Littman, 2005; Turney, 2005; Nastase et al., 2006) have used their class scheme and data set. Moldovan et al. (2004) propose a 35-class scheme to classify relations in various phrases; the same scheme has been applied to noun compounds and other noun phrases (Girju et al., 2005). Chklovski and Pantel (2004) introduce a 5-class set, designed specifically for characterizing verb-verb semantic relations. Stephens et al. (2001) propose 17 classes targeted to relations between genes. Lapata (2002) presents a binary classification of relations in nominalizations.

There is little consensus on the relation sets and algorithms for analyzing semantic relations, and it seems unlikely that any single scheme could work for all applications. For example, the gene-gene relation scheme of Stephens et al. (2001), with relations like *X phosphorylates Y*, is unlikely to be transferred easily to general text.

We have created a benchmark data set to allow the evaluation of different semantic relation classification algorithms. We do not presume to propose a single classification scheme, however alluring it would

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 13–18,
Prague, June 2007. ©2007 Association for Computational Linguistics

Relation	Training data		Test data		Agreement	Example
	positive	set size	positive	set size	(independent tagging)	
Cause-Effect	52.1%	140	51.3%	80	86.1%	laugh (cause) wrinkles (effect)
Instrument-Agency	50.7%	140	48.7%	78	69.6%	laser (instrument) printer (agency)
Product-Producer	60.7%	140	66.7%	93	68.5%	honey (product) bee (producer)
Origin-Entity	38.6%	140	44.4%	81	77.8%	message (entity) from outer-space (origin)
Theme-Tool	41.4%	140	40.8%	71	47.8%	news (theme) conference(tool)
Part-Whole	46.4%	140	36.1%	72	73.2%	the door (part) of the car (whole)
Content-Container	46.4%	140	51.4%	74	69.1%	the apples (content) in the basket (container)

Table 1: Data set statistics

be to try to design a unified standard – it would be likely to have shortcomings just as any of the others we have just reviewed. Instead, we have decided to focus on separate semantic relations that many researchers list in their relation sets. We have built annotated data sets for seven such relations. Every data set supports a separate binary classification task.

2 Building the Annotated Data Sets

Ours is a new evaluation task, so we began with data set creation and annotation guidelines. The data set that Nastase and Szpakowicz (2003) created had relation labels *and* part-of-speech and WordNet sense annotations, to facilitate classification. (Moldovan et al., 2004; Girju et al., 2005) gave the annotators an example of each phrase in a sentence along with WordNet senses and position of arguments. Our annotations include all these, to support a variety of methods (since we work with relations between nominals, the part of speech is always *noun*). We have used WordNet 3.0 on the Web and sense index tags.

We chose the following semantic relations: Cause-Effect, Content-Container, Instrument-Agency, Origin-Entity, Part-Whole, Product-Producer and Theme-Tool. We wrote seven detailed definitions, including restrictions and conventions, plus prototypical positive and near-miss negative examples. For each relation separately, we based data collection on wild-card search patterns that Google allows. We built the patterns manually, following Hearst (1992) and Nakov and Hearst (2006). Instances of the relation Content-Container, for example, come up in response to queries such as "* contains *", "* holds *", "the * in the *". Following the model of the Senseval-3 English Lexical Sample Task, we set out to collect 140 training and at least 70 test examples per relation, so we had a number of different patterns to ensure variety. We also aimed to collect a balanced number of positive and negative examples. The use of heuristic patterns to search for both positive and negative examples should naturally result in negative examples that are near misses. We believe that near misses are more useful for supervised learning than negative examples that are generated randomly.

"Among the contents of the <e1>vessel</e1> were a set of carpenter's <e2>tools</e2>, several large storage jars, ceramic utensils, ropes and remnants of food, as well as a heavy load of ballast stones."

WordNet(e1) = "vessel%1:06:00::",
WordNet(e2) = "tool%1:06:00::",
Content-Container(e2, e1) = "true",
Query = "contents of the * were a"

Figure 1: Annotations illustrated

Figure 1 illustrates the annotations. We tag the nominals, so parsing or chunking is not necessary. For Task 4, we define a nominal as a noun or base noun phrase, excluding names entities. A base noun phrase, e.g., *lawn* or *lawn mower*, is a noun with premodifiers. We also exclude complex noun phrases (e.g., with attached prepositional phrases – *the engine of the lawn mower*).

The procedure was the same for each relation. One person gathered the sample sentences (aiming approximately for a similar number of positive and negative examples) and tagged the entities; two other people annotated the sentences with WordNet senses and classified the relations. The detailed relation definitions and the preliminary discussions of positive and negative examples served to maximize the agreement between the annotators. They first classified the data independently, then discussed every disagreement and looked for consensus. Only the agreed-upon examples went into the data sets. Next, we split each data set into 140 training and no fewer than 70 test examples. (We published the training set for the Content-Container relation as development data two months before the test set.) Table 1 shows the number of positive and negative ex-

amples for each relation.[1]

The average inter-annotator agreement on relations (true/false) after the independent annotation step was 70.3%, and the average agreement on WordNet sense labels was 71.9%. In the process of arriving at a consensus between annotators, the definition of each relation was revised to cover explicitly cases where there had been disagreement. We expect that these revised definitions would lead to much higher levels of agreement than the original definitions did.

3 The Participants

The task of classifying semantic relations between nominals has attracted the participation of 14 teams who submitted 15 systems. Table 4 lists the systems, the authors and their affiliations, and brief descriptions. The systems' performance information in terms of precision, recall, F-measure and accuracy, macroaveraged over all relations, appears in Table 3. We computed these measures as described in Lewis (1991).

We distinguish four categories of systems based on the type of information used – WordNet senses and/or Google queries:

A – WordNet = NO & Query = NO;
B – WordNet = YES & Query = NO;
C – WordNet = NO & Query = YES;
D – WordNet = YES & Query = YES.

WordNet = "YES" or WordNet = "NO" tells us only whether a system uses the WordNet sense labels in the data sets. A system may use WordNet internally for varied purposes, but ignore our sense labels; such a system would be in category A or C. Based on the input variation, each submitted system may have up to 4 variations – A,B,C,D.

Table 2 presents three baselines for a relation. *Majority* always guesses either "true" or "false", whichever is the majority in the test set (maximizes accuracy). *Alltrue* always guesses "true" (maximizes recall). *Probmatch* randomly guesses "true" ("false") with the probability matching the distribution of "true" ("false") in the test dataset (balances precision and recall).

We present the results in Table 3 grouped by category, to facilitate system comparison.

[1] As this paper serves also as a documentation of the data set, the order of relations in the table is the same as in the data set.

Type	P	R	F	Acc
majority	81.3	42.9	30.8	57.0
alltrue	48.5	100.0	64.8	48.5
probmatch	48.5	48.5	48.5	51.7

Table 2: Baselines: precision, recall, F-measure and accuracy averaged over the 7 binary classifications.

Team	P	R	F	Acc
A – WordNet = NO & Query = NO				
UCD-FC	66.1	66.7	64.8	66.0
ILK	60.5	69.5	63.8	63.5
UCB†	62.7	63.0	62.7	65.4
UMELB-B	61.5	55.7	57.8	62.7
UTH	56.1	57.1	55.9	58.8
UC3M	48.2	40.3	43.1	49.9
avg±stdev	59.2±6.3	58.7±10.5	58.0±8.1	61.1±6.0
B – WordNet = YES & Query = NO				
UIUC†	79.7	69.8	72.4	76.3
FBK-IRST	70.9	73.4	71.8	72.9
ILK	72.8	70.6	71.5	73.2
UCD-S1	69.9	64.6	66.8	71.4
UCD-PN	62.0	71.7	65.4	67.0
UC3M	66.7	62.8	64.3	67.2
CMU-AT	55.7	66.7	60.4	59.1
UCD-FC	66.4	58.1	60.3	63.6
UMELB-A	61.7	56.8	58.7	62.5
UVAVU	56.8	56.3	56.1	57.7
LCC-SRN	55.9	57.8	51.4	53.7
avg ± stdev	65.3±7.7	64.4±6.5	63.6±6.9	65.9±7.2
C – WordNet = NO & Query = YES				
UCB†	64.2	66.5	65.1	67.0
UCD-FC	66.1	66.7	64.8	66.0
UC3M	49.4	43.9	45.3	50.1
avg±stdev	59.9±9.1	59.0±13.1	58.4±11.3	61.0±9.5
D – WordNet = YES & Query = YES				
UTD-HLT-CG	67.3	65.3	62.6	67.2
UCD-FC	66.4	58.1	60.3	63.6
UC3M	60.9	57.8	58.8	62.3
avg±stdev	64.9±3.5	60.4±4.2	60.6±1.9	64.4±2.5

Systems tagged with † have a Task 4 organizer as part of the team.

Table 3: System performance grouped by category. Precision, recall, F-measure and accuracy macroaveraged over each system's performance on all 7 relations.

4 Discussion

The highest average accuracy on Task 4 was 76.3%. Therefore, the average initial agreement between annotators (70.3%), before revising the definitions, is not an upper bound on the accuracy that can be achieved. That the initial agreement between annotators is not a good indicator of the accuracy that can be achieved is also supported by the low correlation

System	Institution	Team	Description	System Type
UVAVU	Univ. of Amsterdam TNO Science & Industry Free Univ. Amsterdam	Sophia Katrenko Willem Robert van Hage	similarity measures in WordNet; syntactic dependencies; lexical patterns; logical combination of attributes	B
CMU -AT	Carnegie Mellon Univ.	Alicia Tribble Scott E. Fahlman	WordNet; manually-built ontologies; Scone Knowledge Representation Language; semantic distance	B
ILK	Tilburg University	Caroline Sporleder Roser Morante Antal van den Bosch	semantic clusters based on noun similarity; WordNet supersenses; grammatical relation between entities; head of sentence; WEKA	A, B
FBK-IRST	Fondazione Bruno Kessler - IRST	Claudio Giuliano Alberto Lavelli Daniele Pighin Lorenza Romano	shallow and deep syntactic information; WordNet synsets and hypernyms; kernel methods; SVM	B
LCC-SRN	Language Computer Corp.	Adriana Badulescu	named entity recognition; lexical, semantic, syntactic features; decision tree and semantic scattering	B
UMELB-A	Univ. of Melbourne	Su Kim Timothy Baldwin	sense collocations; similarity of constituents; extending training and testing data using similar words	B
UMELB-B	Univ. of Melbourne	Su Kim Timothy Baldwin	similarity of nearest-neighbor matching over the union of senses for the two nominals; cascaded tagging with decreasing thresholds	A
UCB†	Univ. of California at Berkeley	Preslav Nakov Marti Hearst	VSM; joining terms; KNN-1	A, C
UC3M	Univ. Carlos III of Madrid	Isabel Segura Bedmar Doaa Sammy José Luis Martínez Fernández	WordNet path; syntactic features; SVM	A, B, C, D
UCD-S1	Univ. College Dublin	Cristina Butnariu Tony Veale	lexical-semantic categories from WordNet; syntactic patterns from corpora, SVM	B
UCD-FC	Univ. College Dublin	Fintan Costello	WordNet; additional noun compounds tagged corpus; Naive Bayes	A, B, C, D
UCD-PN	Univ. College Dublin	Paul Nulty	WordNet supersenses; web-based frequency counts for specific joining terms; WEKA (SMO)	B
UIUC†	Univ. of Illinois at Urbana Champaign	Roxana Girju Brandon Beamer Suma Bhat Brant Chee Andrew Fister Alla Rozovskaya	features based on WordNet, NomLex-PLUS, grammatical roles, lexico-syntactic patterns, semantic parses	B
UTD-HLT-CG	Univ. of Texas at Dallas	Cristina Nicolae Garbiel Nicolae Sanda Harabagiu	lexico-semantic features from WordNet, VerbNet; semantic features from a PropBank parser; dependency features	D
UTH	Univ. of Tokio	Eiji Aramaki Takeshi Imai Kengo Miyo Kazuhiko Ohe	joining phrases; physical size for entities; web-mining; SVM	A

Systems tagged with † have a Task 4 organizer as part of the team.

Table 4: Short description of the teams and the participating systems.

Relation	Team	Type	P	R	F	Acc	Test size	Base-F	Base-Acc	Avg. rank
Cause-Effect	UIUC	B_4	69.5	100.0	82.0	77.5	80	67.8	51.2	3.4
Instrument-Agency	FBK-IRST	B_4	76.9	78.9	77.9	78.2	78	65.5	51.3	3.4
Product-Producer	UCD-S1	B_4	80.6	87.1	83.7	77.4	93	80.0	66.7	1.7
Origin-Entity	ILK	B_3	70.6	66.7	68.6	72.8	81	61.5	55.6	6.0
Theme-Tool	ILK	B_4	69.0	69.0	69.0	74.6	71	58.0	59.2	6.0
Part-Whole	UC3M	B_4	72.4	80.8	76.4	81.9	72	53.1	63.9	4.5
Content-Container	UIUC	B_4	93.1	71.1	80.6	82.4	74	67.9	51.4	3.1

Table 5: The best results per relation. Precision, recall, F-measure and accuracy macro-averaged over each system's performance on all 7 relations. Base-F shows the baseline F-measure (alltrue), Base-Acc – the baseline accuracy score (majority). The last column shows the average rank for each relation.

of 0.15 between the Acc column in Table 5 and the Agreement column in Table 1.

We performed various analyses of the results, which we summarize here in four questions. We write X_i to refer to four possible system categories (A_i, B_i, C_i, and D_i) with four possible amounts of training data (X_1 for training examples 1 to 35, X_2 for 1 to 70, X_3 for 1 to 105, and X_4 for 1 to 140).

Does more training data help?
Overall, the results suggest that more training data improves the performance. There were 17 cases in which we had results for all four possible amounts of training data. All average F-measure differences, $F(X_4)$–$F(X_i)$ where $X = A$ to D, $i = 1$ to 3, for these 17 sets of results are statistically significant:

$F(X_4)$–$F(X_1)$: N = 17, avg = 8.3, std = 5.8, min = 1.1, max = 19.6, t-value = -5.9, p-value = 0.00001.

$F(X_4)$–$F(X_2)$: N = 17, avg = 4.0, std = 3.7, min = -3.5, max = 10.5, t-value = 4.5, p-value = 0.0002.

$F(X_4)$–$F(X_3)$: N = 17, avg = 0.9, std = 1.7, min = -2.6, max = 4.7, t-value = 2.1, p-value = 0.03.

Does WordNet help?
The statistics show that WordNet is important, although the contribution varies across systems. Three teams submitted altogether 12 results both for A_1–A_4 and B_1–B_4. The average F-measure difference, $F(B_i)$–$F(A_i)$, $i = 1$ to 4, is significant:

$F(B_i)$–$F(A_i)$: N = 12, avg = 6.1, std = 8.4, min = -4.5, max = 21.2, t-value = -2.5, p-value = 0.01.

The results of the UCD-FC system actually went down when WordNet was used. The statistics for the remaining two teams, however, are a bit better:

$F(B_i)$–$F(A_i)$: N = 8, avg = 10.4, std = 6.7, min = -1.0, max = 21.2, t-value = -4.4, p-value = 0.002.

Does knowing the query help?
Overall, knowing the query did not seem to improve the results. Three teams submitted 12 results both for A_1–A_4 and C_1–C_4. The average F-measure difference, $F(C_i)$–$F(A_i)$, $i = 1$ to 4, is not significant:

$F(C_i)$–$F(A_i)$: N = 12, avg = 0.9, std = 1.8, min = -2.0, max = 5.0, t-value = -1.6, p-value = 0.06.

Again, the UCD-FC system differed from the other systems in that the A and C scores were identical, but even averaging over the remaining two systems and 8 cases does not show a statistically significant advantage:

$F(C_i)$–$F(A_i)$: N = 8, avg = 1.3, std = 2.2, min = -2.0, max = 5.0, t-value = -1.7, p-value = 0.07.

Are some relations harder to classify?
Table 5 shows the best results for each relation in terms of precision, recall, and F-measure, per team and system category. Column *Base-F* presents the baseline F-measure (alltrue), while *Base-Acc* the baseline accuracy score (majority). For all seven relations, the best team significantly outperforms the baseline. The category of the best-scoring system in almost every case is B_4 (only the ILK B_4 system scored second on the Origin-Entity relation).

Table 5 suggests that some relations are more difficult to classify than others. The best F-measure ranges from 83.7 for *Product–Producer* to 68.6 for *Origin–Entity*. The difference between the best F-measure and the baseline F-measure ranges from 23.3 for *Part-Whole* to 3.7 for *Product-Producer*. The difference between the best accuracy and the baseline accuracy ranges from 31.0 for *Content-Container* to 10.7 for *Product-Producer*.

The *F* column shows the best result for each relation, but similar differences among the relations may be observed when all results are pooled. The *Avg. rank* column computes the average rank of each relation in the ordered list of relations generated by each system. For example, *Product–Producer* is often listed as the first or the second easiest relation (with an average rank of 1.7), while *Origin–Entity* and *Theme–Tool* are identified as the most difficult

relations to classify (with average ranks of 6.0).

5 Conclusion

This paper describes a new semantic evaluation task, *Classification of Semantic Relations between Nominals*. We have accomplished our goal of providing a framework and a benchmark data set to allow for comparisons of methods for this task. The data included different types of information – lexical semantic information, context, query used – meant to facilitate the analysis of useful sources of information for determining the semantic relation between nominals. The results that the participating systems have reported show successful approaches to this difficult task, and the advantages of using lexical semantic information.

The success of the task – measured in the interest of the community and the results of the participating systems – shows that the framework and the data are useful resources. By making this collection freely accessible, we encourage further research into this domain and integration of semantic relation algorithms in high-end applications.

Acknowledgments

We thank Eneko Agirre, Lluís Màrquez and Richard Wicentowski, the organizers of SemEval 2007, for their guidance and prompt support in all organizational matters. We thank Marti Hearst for valuable advice throughout the task description and debates on semantic relation definitions. We thank the anonymous reviewers for their helpful comments.

References

T. Chklovski and P. Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proc. Conf. on Empirical Methods in Natural Language Processing, EMNLP-04*, pages 33–40, Barcelona, Spain.

R. Girju, D. Moldovan, M. Tatu, and D. Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, 19:479–496.

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. 14th International Conf. on Computational Linguistics (COLING-92)*, pages 539–545.

M. Lapata. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

D.D. Lewis. 1991. Evaluating text categorization. In *Proceedings of the Speech and Natural Language Workshop*, pages 312–318, Asilomar.

D. Moldovan, A. Badulescu, M. Tatu, D. Antohe, and R. Girju. 2004. Models for the semantic classification of noun phrases. In *Proc. Computational Lexical Semantics Workshop at HLT-NAACL 2004*, pages 60–67, Boston, MA.

P. Nakov and M. Hearst. 2006. Using verbs to characterize noun-noun relations. In *Proc. Twelfth International Conf. in Artificial Intelligence (AIMSA-06)*, pages 233–244, Varna,Bulgaria.

V. Nastase and S. Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 285–301, Tilburg, The Netherlands.

V. Nastase, J. Sayyad-Shirabad, M. Sokolova, and S. Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In *Proc. 21st National Conf. on Artificial Intelligence (AAAI 2006)*, pages 781–787, Boston, MA.

B. Rosario and M. Hearst. 2001. Classifying the semantic relations in noun-compounds via domain-specific lexical hierarchy. In *Proc. 2001 Conf. on Empirical Methods in Natural Language Processing (EMNLP-01)*, pages 82–90.

B. Rosario, M. Hearst, and C. Fillmore. 2002. The descent of hierarchy, and selection in relational semantics. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 417–424, Philadelphia, PA.

M. Stephens, M. Palakal, S. Mukhopadhyay, and R. Raje. 2001. Detecting gene relations from MEDLINE abstracts. In *Proc. Sixth Annual Pacific Symposium on Biocomputing*, pages 483–496.

M. Tatu and D. Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proc. Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 371–378, Vancouver, Canada.

P.D. Turney and M.L. Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3):251–278.

P.D. Turney. 2005. Measuring semantic similarity by latent relational analysis. In *Proc. Nineteenth International Joint Conf. on Artificial Intelligence (IJCAI-05)*, pages 1136–1141, Edinburgh, Scotland.