SemEval-2010 task 8

Conference Paper · January 2009

DOI: 10.3115/1621969.1621986

9 authors, including:

Su Nam Kim

37 PUBLICATIONS **513** CITATIONS

SEE PROFILE

Preslav Nakov

Qatar Computing Research Institute

144 PUBLICATIONS **2,134** CITATIONS

SEE PROFILE

Sebastian Padó

Universität Stuttgart

101 PUBLICATIONS **2,005** CITATIONS

SEE PROFILE

Lorenza Romano

Euregio srl, Italy, Bolzano

25 PUBLICATIONS **512** CITATIONS

SEE PROFILE

SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals

Iris Hendrickx[*], **Su Nam Kim**[†], **Zornitsa Kozareva**[‡], **Preslav Nakov**[§], **Diarmuid Ó Séaghdha**[¶], **Sebastian Padó**[‖], **Marco Pennacchiotti**[**], **Lorenza Romano**[††], **Stan Szpakowicz**[‡‡]

Abstract

SemEval-2 Task 8 focuses on *Multi-way classification of semantic relations between pairs of nominals*. The task was designed to compare different approaches to semantic relation classification and to provide a standard testbed for future research. This paper defines the task, describes the training and test data and the process of their creation, lists the participating systems (10 teams, 28 runs), and discusses their results.

1 Introduction

SemEval-2010 Task 8 focused on *semantic relations between pairs of nominals*. For example, *tea* and *ginseng* are in an ENTITY-ORIGIN relation in "*The cup contained tea from dried ginseng.*". The automatic recognition of semantic relations has many applications, such as information extraction, document summarization, machine translation, or construction of thesauri and semantic networks. It can also facilitate auxiliary tasks such as word sense disambiguation, language modeling, paraphrasing, and recognizing textual entailment.

Our goal was to create a testbed for automatic classification of semantic relations. In developing the task we met several challenges: selecting a suitable set of relations, specifying the annotation procedure, and deciding on the details of the task itself. They are discussed briefly in Section 2; see also Hendrickx et al. (2009), which includes a survey of related work. The direct predecessor of Task 8 was *Classification of semantic relations between nominals*, Task 4 at SemEval-1 (Girju et al., 2009), which had a separate binary-labeled dataset for each of seven relations. We have defined SemEval-2010 Task 8 as a multi-way classification task in which the label for each example must be chosen from the complete set of ten relations and the mapping from nouns to argument slots is not provided in advance. We also provide more data: 10,717 annotated examples, compared to 1,529 in SemEval-1 Task 4.

[*] University of Lisbon, iris@clul.ul.pt
[†] University of Melbourne, snkim@csse.unimelb.edu.au
[‡] Information Sciences Institute/University of Southern California, kozareva@isi.edu
[§] National University of Singapore, nakov@comp.nus.edu.sg
[¶] University of Cambridge, do242@cl.cam.ac.uk
[‖] University of Stuttgart, pado@ims.uni-stuttgart.de
[**] Yahoo! Inc., pennacc@yahoo-inc.com
[††] Fondazione Bruno Kessler, romano@fbk.eu
[‡‡] University of Ottawa *and* Polish Academy of Sciences, szpak@site.uottawa.ca

2 Dataset Creation

2.1 The Inventory of Semantic Relations

We first decided on an inventory of semantic relations. Ideally, it should be exhaustive (enable the description of relations between any pair of nominals) and mutually exclusive (each pair of nominals *in context* should map onto only one relation). The literature, however, suggests that no relation inventory satisfies both needs, and, in practice, some trade-off between them must be accepted.

As a pragmatic compromise, we selected nine relations with coverage sufficiently broad to be of general and practical interest. We aimed at avoiding semantic overlap as much as possible. We included, however, two groups of strongly related relations (ENTITY-ORIGIN / ENTITY-DESTINATION and CONTENT-CONTAINER / COMPONENT-WHOLE / MEMBER-COLLECTION) to assess models' ability to make such fine-grained distinctions. Our inventory is given below. The first four were also used in SemEval-1 Task 4, but the annotation guidelines have been revised, and thus no complete continuity should be assumed.

Cause-Effect (CE). An event or object leads to an effect. Example: *those cancers were caused by radiation exposures*

Instrument-Agency (IA). An agent uses an instrument. Example: *phone operator*

Product-Producer (PP). A producer causes a product to exist. Example: *a factory manufactures suits*

Content-Container (CC). An object is physically stored in a delineated area of space. Example: *a bottle full of honey was weighed*

Entity-Origin (EO). An entity is coming or is derived from an origin (e.g., position or material). Example: *letters from foreign countries*

Entity-Destination (ED). An entity is moving towards a destination. Example: *the boy went to bed*

Component-Whole (CW). An object is a component of a larger whole. Example: *my apartment has a large kitchen*

Member-Collection (MC). A member forms a nonfunctional part of a collection. Example: *there are many trees in the forest*

Message-Topic (MT). A message, written or spoken, is about a topic. Example: *the lecture was about semantics*

2.2 Annotation Guidelines

We defined a set of general annotation guidelines as well as detailed guidelines for each semantic relation. Here, we describe the general guidelines, which delineate the scope of the data to be collected and state general principles relevant to the annotation of all relations.[1]

Our objective is to annotate instances of semantic relations which are true in the sense of holding in the most plausible truth-conditional interpretation of the sentence. This is in the tradition of the Textual Entailment or Information Validation paradigm (Dagan et al., 2009), and in contrast to "aboutness" annotation such as semantic roles (Carreras and Màrquez, 2004) or the BioNLP 2009 task (Kim et al., 2009) where negated relations are also labelled as positive. Similarly, we exclude instances of semantic relations which hold only in speculative or counterfactural scenarios. In practice, this means disallowing annotations within the scope of modals or negations, e.g., "*Smoking may/may not have caused cancer in this case.*"

We accept as relation arguments only noun phrases with common-noun heads. This distinguishes our task from much work in Information Extraction, which tends to focus on specific classes of named entities and on considerably more fine-grained relations than we do. Named entities are a specific category of nominal expressions best dealt with using techniques which do not apply to common nouns. We only mark up the semantic heads of nominals, which usually span a single word, except for lexicalized terms such as *science fiction*.

We also impose a syntactic locality requirement on example candidates, thus excluding instances where the relation arguments occur in separate sentential clauses. Permissible syntactic patterns include simple and relative clauses, compounds, and pre- and post-nominal modification. In addition, we did not annotate examples whose interpretation relied on discourse knowledge, which led to the exclusion of pronouns as arguments. Please see the guidelines for details on other issues, including noun compounds, aspectual phenomena and temporal relations.

2.3 The Annotation Process

The annotation took place in three rounds. First, we manually collected around 1,200 sentences for each relation through pattern-based Web search. In order to ensure a wide variety of example sentences, we used a substantial number of patterns for each relation, typically between one hundred and several hundred. Importantly, in the first round, the relation itself was not annotated: the goal was merely to collect positive and near-miss candidate instances. A rough aim was to have 90% of candidates which instantiate the target relation ("positive instances").

In the second round, the collected candidates for each relation went to two independent annotators for labeling. Since we have a multi-way classification task, the annotators used the full inventory of nine relations plus OTHER. The annotation was made easier by the fact that the cases of overlap were largely systematic, arising from general phenomena like metaphorical use and situations where more than one relation holds. For example, there is a systematic potential overlap between CONTENT-CONTAINER and ENTITY-DESTINATION depending on whether the situation described in the sentence is static or dynamic, e.g., "*When I came, the <e1>apples</e1> were already put in the <e2>basket</e2>.*" is CC(e1, e2), while "*Then, the <e1>apples</e1> were quickly put in the <e2>basket</e2>.*" is ED(e1, e2).

In the third round, the remaining disagreements were resolved, and, if no consensus could be achieved, the examples were removed. Finally, we merged all nine datasets to create a set of 10,717 instances. We released 8,000 for training and kept

[1] The full task guidelines are available at `http://docs.google.com/View?id=dfhkmm46_0f63mfvf7`

the rest for testing.[2]

Table 1 shows some statistics about the dataset. The first column (Freq) shows the absolute and relative frequencies of each relation. The second column (Pos) shows that the average share of positive instances was closer to 75% than to 90%, indicating that the patterns catch a substantial amount of "near-miss" cases. However, this effect varies a lot across relations, causing the non-uniform relation distribution in the dataset (first column).[3] After the second round, we also computed inter-annotator agreement (third column, IAA). Inter-annotator agreement was computed on the sentence level, as the percentage of sentences for which the two annotations were identical. That is, these figures can be interpreted as exact-match accuracies. We do not report Kappa, since chance agreement on preselected candidates is difficult to estimate.[4] IAA is between 60% and 95%, again with large relation-dependent variation. Some of the relations were particularly easy to annotate, notably CONTENT-CONTAINER, which can be resolved through relatively clear criteria, despite the systematic ambiguity mentioned above. ENTITY-ORIGIN was the hardest relation to annotate. We encountered ontological difficulties in defining both Entity (e.g., in contrast to Effect) and Origin (as opposed to Cause). Our numbers are on average around 10% higher than those reported by Girju et al. (2009). This may be a side effect of our data collection method. To gather 1,200 examples in realistic time, we had to seek productive search query patterns, which invited certain homogeneity. For example, many queries for CONTENT-CONTAINER centered on "usual suspect" such as *box* or *suitcase*. Many instances of MEMBER-COLLECTION were collected on the basis of from available lists of collective names.

Relation	Freq	Pos	IAA
Cause-Effect	1331 (12.4%)	91.2%	79.0%
Component-Whole	1253 (11.7%)	84.3%	70.0%
Entity-Destination	1137 (10.6%)	80.1%	75.2%
Entity-Origin	974 (9.1%)	69.2%	58.2%
Product-Producer	948 (8.8%)	66.3%	84.8%
Member-Collection	923 (8.6%)	74.7%	68.2%
Message-Topic	895 (8.4%)	74.4%	72.4%
Content-Container	732 (6.8%)	59.3%	95.8%
Instrument-Agency	660 (6.2%)	60.8%	65.0%
Other	1864 (17.4%)	N/A[4]	N/A[4]
Total	10717 (100%)		

Table 1: Annotation Statistics. Freq: Absolute and relative frequency in the dataset; Pos: percentage of "positive" relation instances in the candidate set; IAA: inter-annotator agreement

3 The Task

The participating systems had to solve the following task: given a sentence and two tagged nominals, predict the relation between those nominals *and* the direction of the relation.

We released a detailed scorer which outputs (1) a confusion matrix, (2) accuracy and coverage, (3) precision (P), recall (R), and F_1-Score for each relation, (4) micro-averaged P, R, F_1, (5) macro-averaged P, R, F_1. For (4) and (5), the calculations ignored the OTHER relation. Our official scoring metric is macro-averaged F_1-Score for (9+1)-way classification, taking directionality into account.

The teams were asked to submit test data predictions for varying fractions of the training data. Specifically, we requested results for the first 1000, 2000, 4000, and 8000 training instances, called TD1 through TD4. TD4 was the full training set.

4 Participants and Results

Table 2 lists the participants and provides a rough overview of the system features. Table 3 shows the results. Unless noted otherwise, all quoted numbers are F_1-Scores.

Overall Ranking and Training Data. We rank the teams by the performance of their best system on TD4, since a per-system ranking would favor teams with many submitted runs. UTD submitted the best system, with a performance of over 82%, more than 4% better than the second-best system. FBK_IRST places second, with 77.62%, a tiny margin ahead of ISI (77.57%). Notably, the ISI system outperforms the FBK_IRST system for TD1 to TD3, where it was second-best. The accuracy numbers for TD4 (Acc TD4) lead to the same overall ranking: micro- versus macro-averaging does not appear to make much difference either. A random baseline gives an uninteresting score of 6%. Our competitive baseline system is a simple Naive Bayes classifier which relies on words in the sentential context only; two systems scored below this baseline.

[2]This set includes 891 examples from SemEval-1 Task 4. We re-annotated them and assigned them as the last examples of our *training* dataset to ensure that the test set was unseen.

[3]To what extent our candidate selection produces a biased sample is a question that we cannot address within this paper.

[4]We do not report Pos or IAA for OTHER, since OTHER is a pseudo-relation that was not annotated in its own right. The numbers would therefore not be comparable to other relations.

System	Institution	Team	Description	Res.	Class.
Baseline	Task organizers		local context of 2 words only		BN
ECNU-SR-1	East China Normal University	Man Lan, Yuan Chen, Zhimin Zhou, Yu Xu	stem, POS, syntactic patterns	S	SVM (multi)
ECNU-SR-2,3			features like ECNU-SR-1, different prob. thresholds		SVM (binary)
ECNU-SR-4			stem, POS, syntactic patterns, hyponymy and meronymy relations	WN, S	SVM (multi)
ECNU-SR-5,6			features like ECNU-SR-4, different prob. thresholds		SVM (binary)
ECNU-SR-7			majority vote of ECNU-1,2,4,5		
FBK_IRST-6C32	Fondazione Bruno Kessler	Claudio Giuliano, Kateryna Tymoshenko	3-word window context features (word form, part of speech, orthography) + Cyc; parameter estimation by optimization on training set	Cyc	SVM
FBK_IRST-12C32			FBK_IRST-6C32 + distance features		
FBK_IRST-12VBC32			FBK_IRST-12C32 + verbs		
FBK_IRST-6CA, -12CA, -12VBCA			features as above, parameter estimation by cross-validation		
FBK_NK-RES1	Fondazione Bruno Kessler	Matteo Negri, Milen Kouylekov	collocations, glosses, semantic relations of nominals + context features	WN	BN
FBK_NK-RES 2,3,4			like FBK_NK-RES1 with different context windows and collocation cutoffs		
ISI	Information Sciences Institute, University of Southern California	Stephen Tratz	features from different resources, a noun compound relation system, and various feature related to capitalization, affixes, closed-class words	WN, RT, G	ME
ISTI-1,2	Istituto di scienca e tecnologie dell'informazione "A. Faedo"	Andrea Esuli, Diego Marcheggiani, Fabrizio Sebastiani	Boosting-based classification. Runs differ in their initialization.	WN	2S
JU	Jadavpur University	Santanu Pal, Partha Pakray, Dipankar Das, Sivaji Bandyopadhyay	Verbs, nouns, and prepositions; seed lists for semantic relations; parse features and NEs	WN, S	CRF
SEKA	Hungarian Academy of Sciences	Eszter Simon, Andras Kornai	Levin and Roget classes, n-grams; other grammatical and formal features	RT, LC	ME
TUD-base	Technische Universität Darmstadt	György Szarvas, Iryna Gurevych	word, POS n-grams, dependency path, distance	S	ME
TUD-wp			TUD-base + ESA semantic relatedness scores	+WP	
TUD-comb			TUD-base + own semantic relatedness scores	+WP,WN	
TUD-comb-threshold			TUD-comb with higher threshold for OTHER		
UNITN	University of Trento	Fabio Celli	punctuation, context words, prepositional patterns, estimation of semantic relation	–	DR
UTD	University of Texas at Dallas	Bryan Rink, Sanda Harabagiu	context wods, hypernyms, POS, dependencies, distance, semantic roles, Levin classes, paraphrases	WN, S, G, PB/NB, LC	SVM, 2S

Table 2: Participants of SemEval-2010 Task 8. Res: Resources used (WN: WordNet data; WP: Wikipedia data; S: syntax; LC: Levin classes; G: Google n-grams, RT: Roget's Thesaurus, PB/NB: PropBank/NomBank). Class: Classification style (ME: Maximum Entropy; BN: Bayes Net; DR: Decision Rules/Trees; CRF: Conditional Random Fields; 2S: two-step classification)

System	TD1	TD2	TD3	TD4	Acc TD4	Rank	Best Cat	Worst Cat-9
Baseline	33.04	42.41	50.89	57.52	50.0	-	MC (75.1)	IA (28.0)
ECNU-SR-1	52.13	56.58	58.16	60.08	57.1	4	CE (79.7)	IA (32.2)
ECNU-SR-2	46.24	47.99	69.83	72.59	67.1		CE (84.4)	IA (52.2)
ECNU-SR-3	39.89	42.29	65.47	68.50	62.0		CE (83.4)	IA (46.5)
ECNU-SR-4	67.95	70.58	72.99	74.82	70.5		CE (84.6)	IA (61.4)
ECNU-SR-5	49.32	50.70	72.63	75.43	70.2		CE (85.1)	IA (60.7)
ECNU-SR-6	42.88	45.54	68.87	72.19	65.8		CE (85.2)	IA (56.7)
ECNU-SR-7	58.67	58.87	72.79	75.21	70.2		CE (86.1)	IA (61.8)
FBK_IRST-6C32	60.19	67.31	71.78	76.81	72.4	2	ED (82.6)	IA (69.4)
FBK_IRST-12C32	60.66	67.91	72.04	76.91	72.4		MC (84.2)	IA (68.8)
FBK_IRST-12VBC32	62.64	69.86	73.19	77.11	72.3		ED (85.9)	PP (68.1)
FBK_IRST-6CA	60.58	67.14	71.63	76.28	71.4		CE (82.3)	IA (67.7)
FBK_IRST-12CA	61.33	67.80	71.65	76.39	71.4		ED (81.8)	IA (67.5)
FBK_IRST-12VBCA	63.61	70.20	73.40	77.62	72.8		ED (86.5)	IA (67.3)
FBK_NK-RES1	55.71*	64.06*	67.80*	68.02	62.1	7	ED (77.6)	IA (52.9)
FBK_NK-RES2	54.27*	63.68*	67.08*	67.48	61.4		ED (77.4)	PP (55.2)
FBK_NK-RES3	54.25*	62.73*	66.11*	66.90	60.5		MC (76.7)	IA (56.3)
FBK_NK-RES4	44.11*	58.85*	63.06*	65.84	59.4		MC (76.1)	IA/PP (58.0)
ISI	66.68	71.01	75.51	77.57	72.7	3	CE (87.6)	IA (61.5)
ISTI-1	50.49*	55.80*	61.14*	68.42	63.2	6	ED (80.7)	PP (53.8)
ISTI-2	50.69*	54.29*	59.77*	66.65	61.5		ED (80.2)	IA (48.9)
JU	41.62*	44.98*	47.81*	52.16	50.2	9	CE (75.6)	IA (27.8)
SEKA	51.81	56.34	61.10	66.33	61.9	8	CE (84.0)	PP (43.7)
TUD-base	50.81	54.61	56.98	60.50	56.1	5	CE (80.7)	IA (31.1)
TUD-wp	55.34	60.90	63.78	68.00	63.5		ED (82.9)	IA (44.1)
TUD-comb	57.84	62.52	66.41	68.88	64.6		CE (83.8)	IA (46.8)
TUD-comb-θ	58.35	62.45	66.86	69.23	65.4		CE (83.4)	IA (46.9)
UNITN	16.57*	18.56*	22.45*	26.67	27.4	10	ED (46.4)	PP (0)
UTD	**73.08**	**77.02**	**79.93**	**82.19**	77.9	1	CE (89.6)	IA (68.5)

Table 3: F_1-Score of all submitted systems on the test dataset as a function of training data: TD1=1000, TD2=2000, TD3=4000, TD4=8000 training examples. Official results are calculated on TD4. The results marked with * were submitted after the deadline. The best-performing run for each participant is *italicized*.

As for the amount of training data, we see a substantial improvement for all systems between TD1 and TD4, with diminishing returns for the transition between TD3 and TD4 for many, but not all, systems. Overall, the differences between systems are smaller for TD4 than they are for TD1. The spread between the top three systems is around 10% at TD1, but below 5% at TD4. Still, there are clear differences in the influence of training data size even among systems with the same overall architecture. Notably, ECNU-SR-4 is the second-best system at TD1 (67.95%), but gains only 7% from the eightfold increase of the size of the training data. At the same time, ECNU-SR-3 improves from less than 40% to almost 69%. The difference between the systems is that ECNU-SR-4 uses a multi-way classifier including the class OTHER, while ECNU-SR-3 uses binary classifiers and assigns OTHER if no other relation was assigned with p>0.5. It appears that these probability estimates for classes are only reliable enough for TD3 and TD4.

The Influence of System Architecture. Almost all systems used either MaxEnt or SVM classifiers, with no clear advantage for either. Similarly, two systems, UTD and ISTI (rank 1 and 6) split the task into two classification steps (relation and direction), but the 2nd- and 3rd-ranked systems do not. The use of a sequence model such as a CRF did not show a benefit either.

The systems use a variety of resources. Generally, richer feature sets lead to better performance (although the differences are often small – compare the different FBK_IRST systems). This improvement can be explained by the need for semantic generalization from training to test data. This need can be addressed using WordNet (contrast ECNU-1 to -3 with ECNU-4 to -6), the Google n-gram collection (see ISI and UTD), or a "deep" semantic resource (FBK_IRST uses Cyc). Yet, most of these resources are also included in the less successful systems, so beneficial integration of knowledge sources into semantic relation classification seems to be difficult.

System Combination. The differences between the systems suggest that it might be possible to achieve improvements by building an ensemble

system. When we combine the top three systems (UTD, FBK_IRST-12VBCA, and ISI) by predicting their majority vote, or OTHER if there was none, we obtain a small improvement over the UTD system with an F_1-Score of 82.79%. A combination of the top five systems using the same method shows a worse performance, however (80.42%). This suggests that the best system outperforms the rest by a margin that cannot be compensated with system combination, at least not with a crude majority vote. We see a similar pattern among the ECNU systems, where the ECNU-SR-7 combination system is outperformed by ECNU-SR-5, presumably since it incorporates the inferior ECNU-SR-1 system.

Relation-specific Analysis. We also analyze the performance on individual relations, especially the extremes. There are very stable patterns across all systems. The best relation (presumably the easiest to classify) is CE, far ahead of ED and MC. Notably, the performance for the best relation is 75% or above for almost all systems, with comparatively small differences between the systems. The hardest relation is generally IA, followed by PP.[5] Here, the spread among the systems is much larger: the highest-ranking systems outperform others on the difficult relations. Recall was the main problem for both IA and PP: many examples of these two relations are misclassified, most frequently as OTHER. Even at TD4, these datasets seem to be less homogeneous than the others. Intriguingly, PP shows a very high inter-annotator agreement (Table 1). Its difficulty may therefore be due not to questionable annotation, but to genuine variability, or at least the selection of difficult patterns by the dataset creator. Conversely, MC, among the easiest relations to model, shows only a modest IAA.

Difficult Instances. There were 152 examples that are classified incorrectly by all systems. We analyze them, looking for sources of errors. In addition to a handful of annotation errors and some borderline cases, they are made up of instances which illustrate the limits of current shallow modeling approaches in that they require more lexical knowledge and complex reasoning. A case in point: *The bottle carrier converts your <e1>bottle</e1> into a <e2>canteen</e2>*. This instance of OTHER is misclassified either as CC (due to the nominals) or as ED (because of the preposition *into*). Another example: *[...] <e1>Rudders</e1> are used by <e2>towboats</e2> and other vessels that require a high degree of manoeuvrability.* This is an instance of CW misclassified as IA, probably on account of the verb *use* which is a frequent indicator of an agentive relation.

[5] The relation OTHER, which we ignore in the overall F_1-score, does even worse, often below 40%. This is to be expected, since the OTHER examples in our datasets are near misses for other relations, thus making a very incoherent class.

5 Discussion and Conclusion

There is little doubt that 19-way classification is a non-trivial challenge. It is even harder when the domain is lexical semantics, with its idiosyncrasies, and when the classes are not necessarily disjoint, despite our best intentions. It speaks to the success of the exercise that the participating systems' performance was generally high, well over an order of magnitude above random guessing. This may be due to the impressive array of tools and lexical-semantic resources deployed by the participants.

Section 4 suggests a few ways of interpreting and analyzing the results. Long-term lessons will undoubtedly emerge from the workshop discussion. One optimistic-pessimistic conclusion concerns the size of the training data. The notable gain TD3 → TD4 suggests that even more data would be helpful, but that is so much easier said than done: it took the organizers well in excess of 1000 person-hours to pin down the problem, hone the guidelines and relation definitions, construct sufficient amounts of trustworthy training data, and run the task.

References

X. Carreras and L. Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proc. CoNLL-04*, Boston, MA.

I. Dagan, B. Dolan, B. Magnini, and D. Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret. 2009. Classification of semantic relations between nominals. *Language Resources and Evaluation*, 43(2):105–121.

I. Hendrickx, S. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. 2009. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proc. NAACL Workshop on Semantic Evaluations*, Boulder, CO.

J. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proc. BioNLP-09*, Boulder, CO.