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

We present a brief overview of the main challenges in the extraction of semantic relations from English text, and discuss the shortcomings of previous data sets and shared tasks. This leads us to introduce a new task, which will be part of SemEval-2010: multi-way classification of mutually exclusive semantic relations between pairs of common nominals. The task is designed to compare different approaches to the problem and to provide a standard testbed for future research, which can benefit many applications in Natural Language Processing.

1 Introduction

The computational linguistics community has a considerable interest in robust knowledge extraction, both as an end in itself and as an intermediate step in a variety of Natural Language Processing (NLP) applications. *Semantic relations between pairs of words* are an interesting case of such semantic knowledge. It can guide the recovery of useful facts about the world, the interpretation of a sentence, or even discourse processing. For example, *pears* and *bowl* are connected in a CONTENT-CONTAINER relation in the sentence "*The bowl contained apples, pears, and oranges.*", while *ginseng* and *taste* are in an ENTITY-ORIGIN relation in "*The taste is not from alcohol, but from the ginseng.*".

The automatic recognition of semantic relations can have many applications, such as information extraction (IE), 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 or recognizing textual entailment. For example, semantic network construction can benefit from detecting a FUNCTION relation between *airplane* and *transportation* in "*the airplane is used for transportation*" or a PART-WHOLE relation in "*the car has an engine*". Similarly, all domains that require deep understanding of text relations can benefit from knowing the relations that describe events like ACQUISITION between named entities in "*Yahoo has made a definitive agreement to acquire Flickr*".

In this paper, we focus on the recognition of *semantic relations between pairs of common nominals*. We present a task which will be part of the SemEval-2010 evaluation exercise and for which we are developing a new benchmark data set. This data set and the associated task address three significant problems encountered in previous work: (1) the definition of a suitable set of relations; (2) the incorporation of context; (3) the desire for a realistic experimental design. We outline these issues in Section 2. Section 3 describes the inventory of relations we adopted for the task. The annotation process, the design of the task itself and the evaluation methodology are presented in Sections 4-6.

[*] University of Antwerp, iris.hendrickx@ua.ac.be

[†] University of Melbourne, snkim@csse.unimelb.edu.au

[‡] University of Alicante, zkozareva@dlsi.ua.es

[§] National University of Singapore, nakov@comp.nus.edu.sg

[¶] University of Cambridge, do242@cl.cam.ac.uk

[||] University of Stuttgart, pado@stanford.edu

[**] 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 Semantic Relation Classification: Issues

2.1 Defining the Relation Inventory

A wide variety of relation classification schemes exist in the literature, reflecting the needs and granularities of various applications. Some researchers only investigate relations between named entities or internal to noun-noun compounds, while others have a more general focus. Some schemes are specific to a domain such as biomedical text.

Rosario and Hearst (2001) classify noun compounds from the domain of medicine into 13 classes that describe the semantic relation between the head noun and the modifier. 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. Stephens et al. (2001) propose 17 very specific classes targeting relations between genes. Nastase and Szpakowicz (2003) address the problem of classifying noun-modifier relations in general text. They propose a two-level hierarchy, with 5 classes at the first level and 30 classes at the second one; other researchers (Kim and Baldwin, 2005; Nakov and Hearst, 2008; Nastase et al., 2006; Turney, 2005; Turney and Littman, 2005) 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). Lapata (2002) presents a binary classification of relations in nominalizations. Pantel and Pennacchiotti (2006) concentrate on five relations in an IE-style setting. In short, there is little agreement on relation inventories.

2.2 The Role of Context

A fundamental question in relation classification is whether the relations between nominals should be considered *out of context* or *in context*. When one looks at real data, it becomes clear that context does indeed play a role. Consider, for example, the noun compound *wood shed*: it may refer either to a shed *made of* wood, or to a shed of any material *used to store* wood. This ambiguity is likely to be resolved in particular contexts. In fact, most NLP applications will want to determine not all possible relations between two words, but rather the relation between two instances in a particular context. While the integration of context is common in the field of IE (cf. work in the context of ACE[1]), much of the existing literature on relation extraction considers word pairs out of context (thus, types rather than tokens). A notable exception is SemEval-2007 Task 4 *Classification of Semantic Relations between Nominals* (Girju et al., 2007; Girju et al., 2008), the first to offer a standard benchmark data set for seven semantic relations between common nouns in context.

2.3 Style of Classification

The design of SemEval-2007 Task 4 had an important limitation. The data set avoided the challenge of defining a single unified standard classification scheme by creating seven separate training and test sets, one for each semantic relation. That made the relation recognition task on each data set a simple *binary* (positive / negative) classification task.[2] Clearly, this does not easily transfer to practical NLP settings, where *any* relation can hold between a pair of nominals which occur in a sentence or a discourse.

2.4 Summary

While there is a substantial amount of work on relation extraction, the lack of standardization makes it difficult to compare different approaches. It is known from other fields that the availability of standard benchmark data sets can provide a boost to the advancement of a field. As a first step, SemEval-2007 Task 4 offered many useful insights into the performance of different approaches to semantic relation classification; it has also motivated follow-up research (Davidov and Rappoport, 2008; Katrenko and Adriaans, 2008; Nakov and Hearst, 2008; Ó Séaghdha and Copestake, 2008).

Our objective is to build on the achievements of SemEval-2007 Task 4 while addressing its shortcomings. In particular, we consider a larger set of semantic relations (9 instead of 7), we assume a proper multi-class classification setting, we emulate the effect of an "open" relation inventory by means of a tenth class OTHER, and we will release to the research community a data set with a considerably

[1] `http://www.itl.nist.gov/iad/mig/tests/ace/`

[2] Although it was not designed for a multi-class set-up, some subsequent publications tried to use the data sets in that manner.

larger number of examples than SemEval-2007 Task 4 or other comparable data sets. The last point is crucial for ensuring the robustness of the performance estimates for competing systems.

3 Designing an Inventory of Semantic Relations Between Nominals

We begin by considering the first of the problems listed above: defining of an inventory of semantic relations. Ideally, it should be exhaustive (should allow 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 no such inventory that could satisfy all needs. In practice, one always must decide on a trade-off between these two properties. For example, the gene-gene relation inventory of Stephens et al. (2001), with relations like *X phosphorylates Y*, arguably allows no overlaps, but is too specific for applications to general text.

On the other hand, schemes aimed at exhaustiveness tend to run into overlap issues, due to such fundamental linguistic phenomena as metaphor (Lakoff, 1987). For example, in the sentence *Dark clouds gather over Nepal.*, the relation between *dark clouds* and *Nepal* is literally a type of ENTITY-DESTINATION, but in fact it refers to the ethnic unrest in Nepal.

We seek a pragmatic compromise between the two extremes. We have selected nine relations with sufficiently broad coverage to be of general and practical interest. We aim at avoiding "real" overlap to the extent that this is possible, but we include two sets of similar relations (ENTITY-ORIGIN/ENTITY-DESTINATION and CONTENT-CONTAINER/COMPONENT-WHOLE/MEMBER-COLLECTION), which can help assess the models' ability to make such fine-grained distinctions.[3]

As in Semeval-2007 Task 4, we give ordered two-word names to the relations, where each word describes the role of the corresponding argument. The full list of our nine relations follows[4] (the definitions we show here are intended to be indicative rather than complete):

Cause-Effect. An event or object leads to an effect. Example: *Smoking causes cancer.*

Instrument-Agency. An agent uses an instrument. Example: *laser printer*

Product-Producer. A producer causes a product to exist. Example: *The farmer grows apples.*

Content-Container. An object is physically stored in a delineated area of space, the container. Example: *Earth is located in the Milky Way.*

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

Entity-Destination. An entity is moving towards a destination. Example: *The boy went to bed.*

Component-Whole. An object is a component of a larger whole. Example: *My apartment has a large kitchen.*

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

Communication-Topic. An act of communication, whether written or spoken, is about a topic. Example: *The lecture was about semantics.*

We add a tenth element to this set, the pseudo-relation OTHER. It stands for any relation which is not one of the nine explicitly annotated relations. This is motivated by modelling considerations. Presumably, the data for OTHER will be very nonhomogeneous. By including it, we force any model of the complete data set to correctly identify the decision boundaries between the individual relations and "everything else". This encourages good generalization behaviour to larger, noisier data sets commonly seen in real-world applications.

3.1 Semantic Relations versus Semantic Roles

There are three main differences between our task (classification of semantic relations between nominals) and the related task of automatic labeling of semantic roles (Gildea and Jurafsky, 2002).

The first difference is to do with the linguistic phenomena described. Lexical resources for theories of semantic roles such as FrameNet (Fillmore et

[3] COMPONENT-WHOLE and MEMBER-COLLECTION are proper subsets of PART-WHOLE, one of the relations in SemEval-2007 Task 4.

[4] We have taken the first five from SemEval-2007 Task 4.

al., 2003) and PropBank (Palmer et al., 2005) have been developed to describe the linguistic realization patterns of events and states. Thus, they target primarily verbs (or event nominalizations) and their dependents, which are typically nouns. In contrast, semantic relations may occur between all parts of speech, although we limit our attention to nominals *in this task*. Also, semantic role descriptions typically relate an event to a set of multiple participants and props, while semantic relations are in practice (although not necessarily) binary.

The second major difference is the syntactic context. Theories of semantic roles usually developed out of syntactic descriptions of verb valencies, and thus they focus on describing the linking patterns of verbs and their direct dependents, phenomena like raising and noninstantiations notwithstanding (Fillmore, 2002). Semantic relations are not tied to predicate-argument structures. They can also be established within noun phrases, noun compounds, or sentences more generally (cf. the examples above).

The third difference is that of the level of generalization. FrameNet currently contains more than 825 different frames (event classes). Since the semantic roles are designed to be interpreted at the frame level, there is *a priori* a very large number of unrelated semantic roles. There is a rudimentary frame hierarchy that defines mappings between roles of individual frames,[5] but it is far from complete. The situation is similar in PropBank. PropBank does use a small number of semantic roles, but these are again to be interpreted at the level of individual predicates, with little cross-predicate generalization. In contrast, all of the semantic relation inventories discussed in Section 1 contain fewer than 50 types of semantic relations. More generally, semantic relation inventories attempt to generalize relations across wide groups of verbs (Chklovski and Pantel, 2004) and include relations that are not verb-centered (Nastase and Szpakowicz, 2003; Moldovan et al., 2004). Using the same labels for similar semantic relations facilitates supervised learning. For example, a model trained with examples of *sell* relations should be able to transfer what it has learned to *give* relations. This has the potential of adding crucial robustness and coverage to analysis tools in NLP applications based on semantic relations.

1. People in Hawaii might be feeling <e1>aftershocks</e1> from that powerful <e2>earthquake</e2> for weeks.

2. My new <e1>apartment</e1> has a <e2>large kitchen</e2>.

Figure 1: Two example sentences with annotation

4 Annotation

The next step in our study will be the actual annotation of relations between nominals. For the purpose of annotation, we define a *nominal* as a noun or a base noun phrase. A base noun phrase is a noun and its pre-modifiers (e.g., nouns, adjectives, determiners). We do not include complex noun phrases (e.g., noun phrases with attached prepositional phrases or relative clauses). For example, *lawn* is a noun, *lawn mower* is a base noun phrase, and *the engine of the lawn mower* is a complex noun phrase.

We focus on heads that are common nouns. This emphasis distinguishes our task from much work in IE, which focuses on named entities and on considerably more fine-grained relations than we do. For example, Patwardhan and Riloff (2007) identify categories like *Terrorist organization* as participants in terror-related semantic relations, which consists predominantly of named entities. We feel that named entities are a specific category of nominal expressions best dealt with using techniques which do not apply to common nouns; for example, they do not lend themselves well to semantic generalization.

Figure 1 shows two examples of annotated sentences. The XML tags <e1> and <e2> mark the target nominals. Since all nine proper semantic relations in this task are asymmetric, the ordering of the two nominals must be taken into account. In example 1, CAUSE-EFFECT(e1, e2) does not hold, although CAUSE-EFFECT(e2, e1) would. In example 2, COMPONENT-WHOLE(e2, e1) holds.

We are currently developing annotation guidelines for each of the relations. They will give a precise definition for each relation and some prototypical examples, similarly to SemEval-2007 Task 4.

The annotation will take place in two rounds. In the first round, we will do a coarse-grained search

[5] For example, it relates the BUYER role of the COMMERCE_SELL frame (verb *sell*) to the RECIPIENT role of the GIVING frame (verb *give*).

for positive examples for each relation. We will collect data from the Web using a semi-automatic, pattern-based search procedure. In order to ensure a wide variety of example sentences, we will use several dozen patterns per relation. We will also ensure that patterns retrieve both positive and negative example sentences; the latter will help populate the OTHER relation with realistic *near-miss* negative examples of the other relations. The patterns will be manually constructed following the approach of Hearst (1992) and Nakov and Hearst (2008).[6]

The example collection for each relation R will be passed to two independent annotators. In order to maintain exclusivity of relations, only examples that are negative for all relations but R will be included as positive and only examples that are negative for all nine relations will be included as OTHER. Next, the annotators will compare their decisions and assess inter-annotator agreement. Consensus will be sought; if the annotators cannot agree on an example it will not be included in the data set, but it will be recorded for future analysis.

Finally, two other task organizers will look for overlap across all relations. They will discard any example marked as positive in two or more relations, as well as examples in OTHER marked as positive in any of the other classes. The OTHER relation will, then, consist of examples that are negatives for all other relations and near-misses for any relation.

Data sets. The annotated data will be divided into a training set, a development set and a test set. There will be 1000 annotated examples for each of the ten relations: 700 for training, 100 for development and 200 for testing. All data will be released under the *Creative Commons Attribution 3.0 Unported License*[7]. The annotation guidelines will be included in the distribution.

5 The Classification Task

The actual task that we will run at SemEval-2010 will be a multi-way classification task. Not all pairs of nominals in each sentence will be labeled, so the gold-standard boundaries of the nominals to be classified will be provided as part of the test data.

In contrast with Semeval 2007 Task 4, in which the ordering of the entities was provided with each example, we aim at a more realistic scenario in which the ordering of the labels is not given. Participants in the task will be asked to discover both the relation and the order of the arguments. Thus, the more challenging task is to identify the *most informative ordering and relation* between a pair of nominals. The stipulation "most informative" is necessary since with our current set of asymmetrical relations that includes OTHER, each pair of nominals that instantiates a relation in one direction (e.g., REL(e1, e2)), instantiates OTHER in the inverse direction (OTHER (e2, e1)). Thus, the correct answers for the two examples in Figure 1 are CAUSE-EFFECT (earthquake, aftershocks) and COMPONENT-WHOLE (large kitchen, apartment).

Note that unlike in SemEval-2007 Task 4, we will not provide manually annotated WordNet senses, thus making the task more realistic. WordNet senses did, however, serve for disambiguation purposes in SemEval-2007 Task 4. We will therefore have to assess the effect of this change on inter-annotator agreement.

6 Evaluation Methodology

The official ranking of the participating systems will be based on their macro-averaged *F-scores* for the nine proper relations. We will also compute and report their *accuracy* over all ten relations, including OTHER. We will further analyze the results quantitatively and qualitatively to gauge which relations are most difficult to classify.

Similarly to SemEval-2007 Task 4, in order to assess the effect of varying quantities of training data, we will ask the teams to submit several sets of guesses for the labels for the test data, using varying fractions of the training data. We may, for example, request test results when training on the first 50, 100, 200, 400 and all 700 examples from each relation.

We will provide a Perl-based automatic evaluation tool that the participants can use when training/tuning/testing their systems. We will use the same tool for the official evaluation.

7 Conclusion

We have introduced a new task, which will be part of SemEval-2010: multi-way classification of semantic

[6] Note that, unlike in Semeval 2007 Task 4, we will not release the patterns to the participants.

[7] `http://creativecommons.org/licenses/by/3.0/`

relations between pairs of common nominals. The task will compare different approaches to the problem and provide a standard testbed for future research, which can benefit many NLP applications.

The description we have presented here should be considered preliminary. We invite the interested reader to visit the official task website `http://semeval2.fbk.eu/semeval2.php?location=tasks\#T11`, where up-to-date information will be published; there is also a discussion group and a mailing list.

References

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proc. EMNLP 2004*, pages 33–40.

Dmitry Davidov and Ari Rappoport. 2008. Classification of semantic relationships between nominals using pattern clusters. In *Proc. ACL-08: HLT*, pages 227–235.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

Charles J. Fillmore. 2002. FrameNet and the linking between semantic and syntactic relations. In *Proc. COLING 2002*, pages 28–36.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Roxana Girju, Dan Moldovan, Marta Tatu, , and Dan Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, 19:479–496.

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 *Proc. 4th Semantic Evaluation Workshop (SemEval-2007)*.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2008. Classification of semantic relations between nominals. *Language Resources and Evaluation*. In print.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. COLING 92*, pages 539–545.

Sophia Katrenko and Pieter Adriaans. 2008. Semantic types of some generic relation arguments: Detection and evaluation. In *Proc. ACL-08: HLT, Short Papers*, pages 185–188.

Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using WordNet similarity. In *Proc. IJCAI*, pages 945–956.

George Lakoff. 1987. *Women, fire, and dangerous things*. University of Chicago Press, Chicago, IL.

Maria Lapata. 2002. The disambiguation of nominalisations. *Computational Linguistics*, 28:357–388.

Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 60–67.

Preslav Nakov and Marti A. Hearst. 2008. Solving relational similarity problems using the web as a corpus. In *Proc. ACL-08: HLT*, pages 452–460.

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

Vivi Nastase, Jelber Sayyad-Shirabad, Marina Sokolova, and Stan Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In *Proc. AAAI*, pages 781–787.

Diarmuid Ó Séaghdha and Ann Copestake. 2008. Semantic classification with distributional kernels. In *Proc. COLING 2008*, pages 649–656.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proc. COLING/ACL*, pages 113–120.

Siddharth Patwardhan and Ellen Riloff. 2007. Effective information extraction with semantic affinity patterns and relevant regions. In *Proc. EMNLP-CoNLL)*, pages 717–727.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proc. EMNLP 2001*, pages 82–90.

Barbara Rosario, Marti Hearst, and Charles Fillmore. 2002. The descent of hierarchy, and selection in relational semantics. In *Proc. ACL-02*, pages 247–254.

Matthew Stephens, Mathew Palakal, Snehasis Mukhopadhyay, Rajeev Raje, and Javed Mostafa. 2001. Detecting gene relations from Medline abstracts. In *Pacific Symposium on Biocomputing*, pages 483–495.

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

Peter D. Turney. 2005. Measuring semantic similarity by latent relational analysis. In *Proc. IJCAI*, pages 1136–1141.