SemEval-2010 Task 7: Argument Selection and Coercion

James Pustejovsky
Computer Science Department
Brandeis University
Waltham, Massachusetts, USA
jamesp@cs.brandeis.edu

Anna Rumshisky
Computer Science Department
Brandeis University
Waltham, Massachusetts, USA
arum@cs.brandeis.edu

Abstract

In this paper, we describe the *Argument Selection and Coercion* task, currently in development for the SemEval-2 evaluation exercise scheduled for 2010. This task involves characterizing the type of compositional operation that exists between a predicate and the arguments it selects. Specifically, the goal is to identify whether the type that a verb selects is satisfied directly by the argument, or whether the argument must change type to satisfy the verb typing. We discuss the problem in detail and describe the data preparation for the task.

1 Introduction

In recent years, a number of annotation schemes that encode semantic information have been developed and used to produce data sets for training machine learning algorithms. Semantic markup schemes that have focused on annotating entity types and, more generally, word senses, have been extended to include semantic relationships between sentence elements, such as the semantic role (or label) assigned to the argument by the predicate (Palmer et al., 2005; Ruppenhofer et al., 2006; Kipper, 2005; Burchardt et al., 2006; Ohara, 2008; Subirats, 2004).

In this task, we take this one step further, in that this task attempts to capture the "compositional history" of the argument selection relative to the predicate. In particular, this task attempts to identify the operations of type adjustment induced by a predicate over its arguments when they do not match its selectional properties. The task is defined as follows: for each argument of a predicate, identify whether the entity in that argument position satisfies the type expected by the predicate. If not, then one needs to identify how the entity in that position satisfies the typing expected by the predicate; that is, to identify the source and target types in a type-shifting (or coercion) operation.

Consider the example below, where the verb *report* normally selects for a human in subject position as in (1). Notice, however, that through a metonymic interpretation, this constraint can be violated as demonstrated in (1).

(1) a. John reported in late from Washington.
b. Washington reported in late.

Neither the surface annotation of entity extents and types, nor assigning semantic roles associated with the predicate would reflect in this case a crucial point: namely, that in order for the typing requirements of the predicate to be satisfied, what has been referred to a *type coercion* or a *metonymy* (Hobbs et al., 1993; Pustejovsky, 1991; Nunberg, 1979; Egg, 2005) has taken place.

The SemEval Metonymy task (Markert and Nissim, 2007) was a good attempt to annotate such metonymic relations over a larger data set. This task involved two types with their metonymic variants:

(2) i. **Categories for Locations**: literal, place-for-people, place-for-event, place-for-product;
ii. **Categories for Organizations**: literal, organization-for-members, organization-for-event, organization-for-product, organization-for-facility.

One of the limitations of this approach, however, is that, while appropriate for these specialized metonymy relations, the annotation specification and resulting corpus are not an informative guide for extending the annotation of argument selection more broadly.

In fact, the metonymy example in (1) is an instance of a much more pervasive phenomenon of type shifting and coercion in argument selection. For example, in (3) below, the sense annotation for the verb *enjoy* should arguably assign similar values to both (3a) and (3b).

(3) a. Mary enjoyed drinking her beer .
b. Mary enjoyed her beer.

The consequence of this, however, is that, under current sense and role annotation strategies, the mapping to a syntactic realization for a given sense is made more complex, and is in fact, perplexing for a clustering or learning algorithm operating over subcategorization types for the verb.

2 Methodology of Annotation

Before introducing the specifics of the argument selection and coercion task, let us review briefly our assumptions regarding the role of annotation within the development and deployment of computational linguistic systems.

We assume that the features we use for encoding a specific linguistic phenomenon are rich enough to capture the desired behavior. These linguistic descriptions are typically distilled from extensive theoretical modeling of the phenomenon. The descriptions in turn form the basis for the annotation values of the specification language, which are themselves the features used in a development cycle for training and testing an identification or labeling algorithm over text. Finally, based on an analysis and evaluation of the performance of a system, the model of the phenomenon may be revised, for retraining and testing.

We call this particular cycle of development the MATTER methodology:

(4) a. **Model**: Structural descriptions provide theoretically-informed attributes derived from empirical observations over the data;

b. **Annotate**: Annotation scheme assumes a feature set that encodes specific structural descriptions and properties of the input data;

c. **Train**: Algorithm is trained over a corpus annotated with the target feature set;

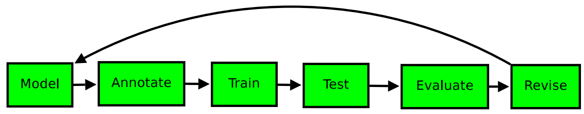

Figure 1: The MATTER Methodology

d. **Test**: Algorithm is tested against held-out data;

e. **Evaluate**: Standardized evaluation of results;

f. **Revise**: Revisit the model, annotation specification, or algorithm, in order to make the annotation more robust and reliable.

Some of the current and completed annotation efforts that have undergone such a development cycle include:

- PropBank (Palmer et al., 2005)
- NomBank (Meyers et al., 2004)
- TimeBank (Pustejovsky et al., 2005)
- Opinion Corpus (Wiebe et al., 2005)
- Penn Discourse TreeBank (Miltsakaki et al., 2004)

3 Task Description

This task involves identifying the selectional mechanism used by the predicate over a particular argument.[1] For the purposes of this task, the possible relations between the predicate and a given argument are restricted to *selection* and *coercion*. In *selection*, the argument NP satisfies the typing requirements of the predicate, as in (5).

(5) a. The spokesman denied the statement (PROPOSITION).
b. The child threw the stone (PHYSICAL OBJECT).
c. The audience didn't believe the rumor (PROPOSITION).

Coercion encompasses all cases when a type-shifting operation must be performed on the complement NP in order to satisfy selectional requirements of the predicate, as in (6). Note that coercion operations may apply to any argument position in a sentence, including the subject, as seen in (6b). Coercion can also be seen as an object of a proposition as in (6c).

(6) a. The president denied the attack (EVENT → PROPOSITION).
b. The White House (LOCATION → HUMAN) denied this statement.
c. The Boston office called with an update (EVENT → INFO).

[1] This task is part of a larger effort to annotate text with compositional operations (Pustejovsky et al., 2009).

The definition of *coercion* will be extended to include instances of type-shifting due to what we term the *qua*-relation.

(7) a. You can crush the pill (PHYSICAL OBJECT) between two spoons. (*Selection*)
b. It is always possible to crush imagination (ABSTRACT ENTITY *qua* PHYSICAL OBJECT) under the weight of numbers. (*Coercion/qua-relation*)

In order to determine whether type-shifting has taken place, the classification task must then involve the following (1) identifying the verb sense and the associated syntactic frame, (2) identifying selectional requirements imposed by that verb sense on the target argument, and (3) identifying semantic type of the target argument. Sense inventories for the verbs and the type templates associated with different syntactic frames will be provided to the participants.

3.1 Semantic Types

In the present task, we use a subset of semantic types from the Brandeis Shallow Ontology (BSO), which is a shallow hierarchy of types developed as a part of the CPA effort (Hanks, 2009; Pustejovsky et al., 2004; Rumshisky et al., 2006). The BSO types were selected for their prevalence in manually identified selection context patterns developed for several hundreds English verbs. That is, they capture common semantic distinctions associated with the selectional properties of many verbs.

The following list of types is currently being used for annotation:

(8) HUMAN, ANIMATE, PHYSICAL OBJECT, ARTIFACT, ORGANIZATION, EVENT, PROPOSITION, INFORMATION, SENSATION, LOCATION, TIME PERIOD, ABSTRACT ENTITY, ATTITUDE, EMOTION, PROPERTY, PRIVILEGE, OBLIGATION, RULE

The subset of types chosen for annotation is purposefully shallow, and is not structured in a hierarchy. For example, we include both HUMAN and ANIMATE in the type system along with PHYSICAL OBJECT. While HUMAN is a subtype of both ANIMATE and PHYSICAL OBJECT, the system should simply choose the most relevant type (i.e. HUMAN) and not be concerned with type inheritance. The present set of types may be revised if necessary as the annotation proceeds.

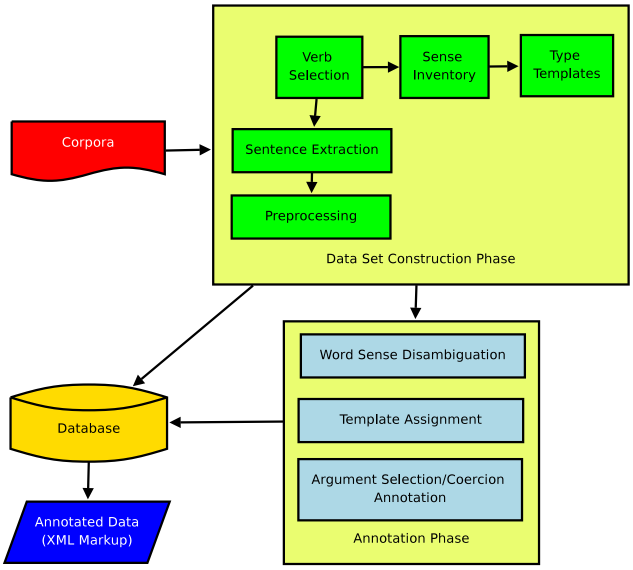

Figure 2: Corpus Development Architecture

4 Resources and Corpus Development

Preparing the data for this task will be done in two phases: *the data set construction phase* and *the annotation phase*. The first phase consists of (1) selecting the target verbs to be annotated and compiling a sense inventory for each target, and (2) data extraction and preprocessing. The prepared data is then loaded into the annotation interface. During the annotation phase, the annotation judgments are entered into the database, and the adjudicator resolves disagreements. The resulting database representation is used by the exporting module to generate the corresponding XML markup or stand-off annotation. The corpus development architecture is shown in Fig. 2.

4.1 Data Set Construction Phase

In the set of target verbs selected for the task, preference will be given to the verbs that are strongly coercive in at least one of their senses, i.e. tend to impose semantic typing on one of their arguments. The verbs will be selected by examining the data from several sources, using the Sketch Engine (Kilgarriff et al., 2004) as described in (Rumshisky and Batiukova, 2008).

An inventory of senses will be compiled for each verb. Whenever possible, the senses will be mapped to OntoNotes (Pradhan et al., 2007) and to the CPA patterns (Hanks, 2009). For each sense, a set of type templates will be compiled, associating each sense with one or more syntactic patterns which will include type specification for all arguments. For example, one of the senses of the verb *deny* is *refuse to grant*. This sense is associated with the following type templates:

(9) HUMAN deny ENTITY to HUMAN
HUMAN deny HUMAN ENTITY

The set of type templates for each verb will be built using a modification of the CPA technique (Hanks and Pustejovsky, 2005; Pustejovsky et al., 2004)).

A set of sentences will be randomly extracted for each target verb from the BNC (BNC, 2000) and the American National Corpus (Ide and Suderman, 2004). This choice of corpora should ensure a more balanced representation of language than is available in commonly annotated WSJ and other newswire text. Each extracted sentence will be automatically parsed, and the sentences organized according to the grammatical relation involving the target verb. Sentences will be excluded from the set if the target argument is expressed as anaphor, or is not present in the sentence. Semantic head for the target grammatical relation will be identified in each case.

4.2 Annotation Phase

Word sense disambiguation will need to be performed as a preliminary stage for the annotation of compositional operations. The annotation task is thus divided into two subtasks, presented successively to the annotator:

(1) Word sense disambiguation of the target predicate

(2) Identification of the compositional relationship between target predicate and its arguments

In the first subtask, the annotator is presented with a set of sentences containing the target verb and the chosen grammatical relation. The annotator is asked to select the most fitting sense of the target verb, or to throw out the example (pick the "N/A" option) if no sense can be chosen either due to insufficient context, because the appropriate sense does not appear in the inventory, or simply no disambiguation can be made in good faith. The interface is shown in Fig. 3. After this step is complete, the appropriate sense is saved into the database, along with the associated type template.

In the second subtask, the annotator is presented with a list of sentences in which the target verb is used in the same sense. The data is annotated one grammatical relation at a time. The annotator is asked to determine whether the argument in the specified grammatical relation to the target belongs to the type associated with that sense in the corresponding template. The illustration of this can be seen in Fig. 4. We will perform double annotation and subsequent adjudication at each of the above annotation stages.

5 Data Format

The test and training data will be provided in XML format. The relation between the predicate (viewed as function) and its argument will be represented by a composition link (`CompLink`) as shown below. In case of *coercion*, there is a mismatch between the source and the target types, and both types need to be identified:

The State Department repeatedly denied the attack.

```
The State Department repeatedly
<SELECTOR sid="s1">denied</SELECTOR>
the
<NOUN nid="n1">attack</NOUN> .
<CompLink cid="cid1" sID="s1"
relatedToNoun="n1" gramRel="dobj"
compType="COERCION"
sourceType="EVENT"
targetType="PROPOSITION"/>
```

When the compositional operation is *selection*, the source and the target types must match:

The State Department repeatedly denied this statement.

```
The State Department repeatedly
<SELECTOR sid="s1">denied</SELECTOR>
this
<NOUN nid="n1">statement</NOUN> .
<CompLink cid="cid1" sID="s1"
relatedToNoun="n1" gramRel="dobj"
compType="selection"
sourceType="PROPOSITION"
targetType="PROPOSITION"/>
```

6 Evaluation Methodology

Precision and recall will be used as evaluation metrics. A scoring program will be supplied for participants. Two subtasks will be evaluated separately:

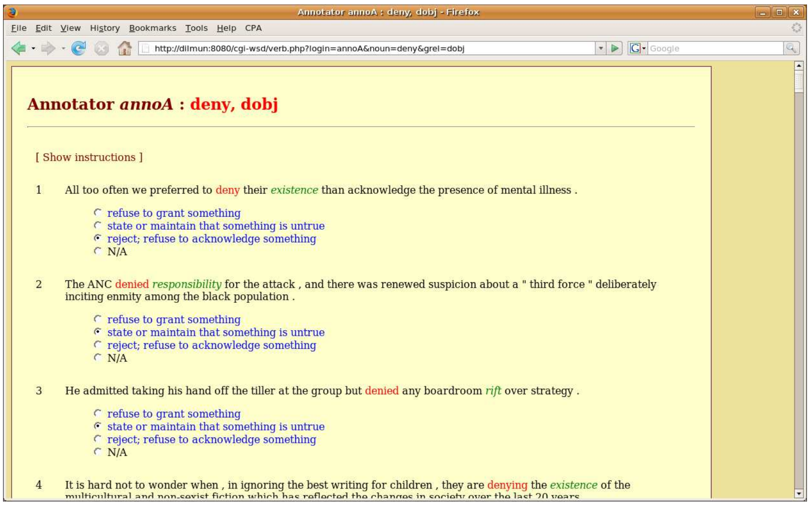

Figure 3: Predicate Sense Disambiguation for *deny*.

(1) identifying the compositional operation (i.e. selection vs. coercion) and (2) identifying the source and target argument type, for each relevant argument. Both subtasks require sense disambiguation which will not be evaluated separately.

Since type-shifting is by its nature a relatively rare event, the distribution between different types of compositional operations in the data set will be necessarily skewed. One of the standard sampling methods for handling class imbalance is downsizing (Japkowicz, 2000; Monard and Batista, 2002), where the number of instances of the major class in the training set is artificially reduced. Another possible alternative is to assign higher error costs to misclassification of minor class instances (Chawla et al., 2004; Domingos, 1999).

7 Conclusion

In this paper, we have described the Argument Selection and Coercion task for SemEval-2, to be held in 2010. This task involves the identifying the relation between a predicate and its argument as one that encodes the compositional history of the selection process. This allows us to distinguish surface forms that directly satisfy the selectional (type) requirements of a predicate from those that are coerced in context. We described some details of a specification language for selection and the annotation task using this specification to identify argument selection behavior. Finally, we discussed data preparation for the task and evaluation techniques for analyzing the results.

References

BNC. 2000. *The British National Corpus*. The BNC Consortium, University of Oxford, *http://www.natcorp.ox.ac.uk/*.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of LREC*, Genoa, Italy.

N. Chawla, N. Japkowicz, and A. Kotcz. 2004. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6.

P. Domingos. 1999. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 155–164. ACM New York, NY, USA.

Marcus Egg. 2005. *Flexible semantics for reinterpretation phenomena*. CSLI, Stanford.

P. Hanks and J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée*.

P. Hanks. 2009. Corpus pattern analysis. CPA Project Page. Retrieved April 11, 2009, from http://nlp.fi.muni.cz/projekty/cpa/.

J. R. Hobbs, M. Stickel, and P. Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

N. Ide and K. Suderman. 2004. The American National Corpus first release. In *Proceedings of LREC 2004*, pages 1681–1684.

Figure 4: Identifying Compositional Relationship for *deny*.

N. Japkowicz. 2000. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, pages 00–05.

A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. *Proceedings of Euralex, Lorient, France*, pages 105–116.

Karin Kipper. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Phd dissertation, University of Pennsylvania, PA.

K. Markert and M. Nissim. 2007. Metonymy resolution at SemEval I: Guidelines for participants. In *Proceedings of the ACL 2007 Conference*.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank project: An interim report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*.

M.C. Monard and G.E. Batista. 2002. Learning with skewed class distributions. *Advances in logic, artificial intelligence and robotics (LAPTEC'02)*.

Geoffrey Nunberg. 1979. The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, 3:143–184.

Kyoko Hirose Ohara. 2008. Lexicon, grammar, and multilinguality in the japanese framenet. In *Proceedings of LREC*, Marrakech, Marocco.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

S. Pradhan, E. Loper, D. Dligach, and M. Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June. Association for Computational Linguistics.

J. Pustejovsky, P. Hanks, and A. Rumshisky. 2004. Automated Induction of Sense in Context. In *COLING 2004, Geneva, Switzerland*, pages 924–931.

J. Pustejovsky, R. Knippen, J. Littman, and R. Sauri. 2005. Temporal and event information in natural language text. *Language Resources and Evaluation*, 39(2):123–164.

J. Pustejovsky, A. Rumshisky, J. Moszkowicz, and O. Batiukova. 2009. GLML: Annotating argument selection and coercion. *IWCS-8: Eighth International Conference on Computational Semantics*.

J. Pustejovsky. 1991. The generative lexicon. *Computational Linguistics*, 17(4).

A. Rumshisky and O. Batiukova. 2008. Polysemy in verbs: systematic relations between senses and their effect on annotation. In *COLING Workshop on Human Judgement in Computational Linguistics (HJCL-2008)*, Manchester, England.

A. Rumshisky, P. Hanks, C. Havasi, and J. Pustejovsky. 2006. Constructing a corpus-based ontology using model bias. In *The 19th International FLAIRS Conference, FLAIRS 2006*, Melbourne Beach, Florida, USA.

J. Ruppenhofer, M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*.

Carlos Subirats. 2004. FrameNet Español. Una red semántica de marcos conceptuales. In *VI International Congress of Hispanic Linguistics*, Leipzig.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.