Guide for Use of Complete Dataset

This document describes the Complete Dataset [1] for Task 4 in SemEval 2007 [2], *Classification of Semantic Relations between Nominals* [3]. If you have any questions about the use of the Complete Dataset that are not answered below or in the description of Task 4, please send a message to the Semantic Relations Google Group [4] or contact any of the authors of *Classification of Semantic Relations between Nominals* [3]. The purpose of this guide is to provide basic information about the Complete Dataset for participants in Task 4. We describe the format of the Complete Dataset, we explain how to use the Complete Dataset, and we suggest some resources that may be useful to participants.

Schedule

The evaluation period will comprise the 5 weeks from **February 26** to **April 1**. During this period, participants can download training and testing data for Task 4 at any time, with the following restrictions:

1. Results for a given task have to be submitted no later than *21 days after downloading the training data* for Task 4.
2. Results for a given task have to be submitted no later than *7 days after downloading the testing data* for Task 4.

See below for more information.

Description of Complete Dataset

The planned release date for the Complete Dataset is February 26, 2007 [5]. The dataset will include data files for the following seven semantic relations:

1. Cause-Effect (e.g., virus-flu)
2. Instrument-Agency (e.g., laser-printer)
3. Product-Producer (e.g., honey-bee)
4. Origin-Entity (e.g., rye-whiskey)
5. Theme-Tool (e.g., soup-pot)
6. Part-Whole (e.g., wheel-car)
7. Content-Container (e.g., apple-basket)

The Complete Dataset will be released in two separate packages:

1. Training Data: a total of 15 files, including 7 training files (140 examples each), 7 definition files (one for each of the 7 semantic relations), and 1 guide file (you are reading the guide file now).
2. Testing Data: a total of 7 files, including 7 testing files (approximately 70 examples each).

Thus there will be a total of 22 files in the Complete Dataset.

This is a training example for the Content-Container relation, which is defined in *Relation 7: Content-Container* [6]:

```
127 "I find it hard to bend and reach and I cannot use the <e1>cupboards</e1> in my
<e2>kitchen</e2>."
WordNet(e1) = "cupboard%1:06:00::", WordNet(e2) = "kitchen%1:06:00::",
Content-Container(e1, e2) = "false", Query = "the * in my kitchen"
Comment: Located-Location or, better, Part-Whole.
```

The first line includes the sentence itself, preceded by a numerical identifier. The two nominals, "cupboards" and "kitchen", are marked by <e1> and <e2> tags. The second line gives the WordNet sense keys for the two nominals [7] and indicates whether the semantic relation between the nominals is a positive ("true") or negative ("false") example of the Content-Container relation [6]. We use WordNet sense keys because, unlike WordNet synset numbers, sense keys are relatively stable across different versions of WordNet. Our preferred version of WordNet is 3.0, but we believe that most of the sense keys for version 2.1 are the same as in version 3.0 [8] (although the synset numbers changed significantly between the two versions). The second line gives the query that was used to find the sentence (mostly by searching on Google). The queries are manually generated heuristic patterns that are intended to find sentences that are examples of the given relation (we aimed for queries that would generate roughly 50% positive examples and 50% negative, although we did not always achieve this aim) [9]. The third line is an optional comment line (not all training examples have a comment line). The comment lines have been added by the annotators, to explain their labeling decisions. The comments are intended for human readers. They should be ignored by the algorithms that participate in the task, and they will not be used in scoring the output of the algorithms.

This is a testing example:

```
127 "I find it hard to bend and reach and I cannot use the <e1>cupboards</e1> in my
<e2>kitchen</e2>."
WordNet(e1) = "cupboard%1:06:00::", WordNet(e2) = "kitchen%1:06:00::",
Content-Container(e1, e2) = "?", Query = "the * in my kitchen"
```

In comparison with the training example, note that the relation, Content-Container(e1, e2), is labeled "?", instead of "true" or "false". For all testing examples, the relations are labeled "?". Also, the comment lines have been removed for all testing examples. After SemEval has finished, the relation labels and the comments will be available.

Note that the order of the entities is important:

```
040 "Your <e1>stomach</e1> is supposed to contain <e2>acid</e2>."
WordNet(e1) = "stomach%1:08:00::", WordNet(e2) = "acid%1:27:00::",
Content-Container(e2, e1) = "true", Query = "to contain acid"
Comment: the best choice, but oddly the definition fails a little (one cannot get rid
of acid).
```

In example 127, we have Content-Container(e1, e2), but we have Content-Container(e2, e1) in example 040. The first term in Content-Container should be the content and the second term should be the container (the name of the relation is intended to act as a reminder for the order of the terms). Example 040 might also have been represented by marking "stomach" with e2 and "acid" with e1, in which case we would have Content-Container(e1, e2), but we decided not to allow this alternative representation. We used the convention that the e1 markup should always precede the e2 markup in the sentence, and then e1 and e2 should be ordered appropriately in the relation (e.g., Content-Container(e2, e1) in example 040).

The above seven semantic relations are not exhaustive; for example, the Hypernym-Hyponym relation is not included. When generating the Complete Dataset, we will consider each relation on its own, as a binary positive-negative classification problem. We will not make any assumptions about whether the relations are overlapping or exclusive. Therefore a positive example of one relation is not necessarily a negative example of another relation. For each relation, approximately half of the sentences will be positive examples and the other half will be near-miss negative examples.

To help motivate Task 4, consider the following potential application. Imagine that we wish to create a new type of search engine for semantic relations. For example, suppose I have just bought a new home, and I am wondering what things I will need to purchase for my new kitchen. I could search for all X such that Content-Container(X, kitchen) = "true". We assume that the search engine will have a predefined set of manually generated heuristic patterns for a few basic semantic relations, such as Content-Container(X, Y). One of the patterns might be "the X in a Y", so that a search for all X such that Content-Container(X, kitchen) = "true" will result in the query "the X in a kitchen". Some of the sentences that are found with this query will be positive examples of Content-Container(X, kitchen) and some will be near-miss negative examples. The challenge of Task 4 is to learn to automatically

distinguish the positive and negative examples. A successful algorithm for this task could be used to filter the query results in a search engine for semantic relations. Other possible applications of a successful algorithm include question answering and paraphrasing.

We released a Trial Dataset on January 3, 2007. The Trial Dataset includes 140 sentences that are positive and negative examples of the Content-Container relation. The Trial Dataset is included in the Training Data in the Complete Dataset; it is one of the seven training files.

Experimenting with the Complete Dataset

Each team may submit 4, 8, 12, or 16 sets of results, as shown in the columns in the following table. The column heading "Amount of training" refers to how much of the training data was used by the algorithm, in order to generate the given results. For example, "Training = 1 to 35" means that the algorithm used the training examples that were numbered from 1 to 35. The column heading "WordNet = NO" indicates that the given results were generated without using the WordNet labels (e.g., WordNet(e1) = "cupboard%1:06:00::"), whereas "WordNet = YES" indicates that the WordNet labels were used. Similarly, the column heading "Query = NO" indicates that the given results were generated without using the Query labels (e.g., Query = "the * in my kitchen"), whereas "Query = YES" indicates that the Query labels were used.

Amount of training	WordNet = NO Query = NO	WordNet = YES Query = NO	WordNet = NO Query = YES	WordNet = YES Query = YES
Training = 1 to 35	A1	B1	C1	D1
Training = 1 to 70	A2	B2	C2	D2
Training = 1 to 105	A3	B3	C3	D3
Training = 1 to 140	A4	B4	C4	D4

A team may submit results for any or all of the four columns in this table (excluding the column "Amount of training"). The table gives a letter-number identifier (e.g., B2) that should be used to identify each set of results. For example, if a team has an algorithm that *must* use WordNet labels and Query labels, they would submit four sets of results, using the identifiers D1, D2, D3, and D4.

Note that the "WordNet = YES" and "WordNet = NO" conditions refer only to whether the algorithm uses (YES) or ignores (NO) the WordNet labels in the datasets. The conditions have nothing to do with whether an algorithm uses WordNet internally for some purpose, such as lemmatization. That is, if an algorithm uses WordNet internally (e.g., for lemmatization or for measuring similarity) but ignores the WordNet labels in the datasets, then it would count as "WordNet = NO".

WordNet contains some Part-Whole information (meronyms and holonyms) and also some Cause-Effect information (for verbs). Particpants are welcome to use this information as a resource in their algorithms. Again, this has nothing to do with the "WordNet = YES" and "WordNet = NO" conditions in the table above.

Performance measures will be calculated automatically by comparing the output of each algorithm to the annotators' labels. The performance of the participants' algorithms will be evaluated based on their success at guessing the hidden true/false labels for the testing sentences. The performance measures will be precision, recall, and F (the harmonic mean of precision and recall). Algorithms will be allowed to skip difficult sentences, for increased precision but decreased recall. The scoring script will accept output in the following format:

```
001 true
002 false
003 skipped
004 skipped
005 false
006 true
```

. . .

For example, the first line of output indicates that the algorithm has guessed that the given relation is true for sentence number 001. The output for a given relation (e.g., Content-Container) under a given experimental condition (e.g., B2 in the table above) should be stored in a file with a name that includes the relation and the experimental condition (e.g., "Content-Container-B2.txt"). Each such file should contain approximately 70 lines, one line for each of the approximately 70 testing examples. The files should then be bundled together and compressed (e.g., tar and gzip). For example, if a team has an algorithm that *must* use WordNet labels and Query labels, their submission would be a bundle of 28 files (4 experimental conditions times 7 relations = 28 files).

The evaluation period will comprise the 5 weeks from February 26 to April 1. During this period, participants can download training and testing data for Task 4 at any time, with the following restrictions:

1. Results for a given task have to be submitted no later than 21 days after downloading the training data for Task 4.
2. Results for a given task have to be submitted no later than 7 days after downloading the testing data for Task 4.

Time constraints will be checked automatically by the downloading application. Before the test period expires, participants will upload the answer files output by their systems, again to the application in the SemEval-2007 website [2]. The SemEval website will centralize all upload and download processes, which will ensure that all participants follow the time constraints and deadlines.

The Complete Dataset (and the Trial Dataset) will be (are) released under the Creative Commons Attribution-Share Alike 2.5 License [10]. There is no need for the participants to fill any license forms in order to access the data. In any work that uses the Complete Dataset, please acknowledge the authors, as follows:

> Roxana Girju, Marti Hearst, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret (2007). Classification of Semantic Relations between Nominals: Dataset for Task 4 in SemEval 2007, *4th International Workshop on Semantic Evaluations*, June 23-24, 2007, Prague, Czech Republic.

Resources

All resources are allowed for Task 4 (e.g., lexicons, corpora, part-of-speech tagging, parsing), but the algorithms must be automated (i.e., no human in the loop). We anticipate that many of the participants will use supervised machine learning algorithms to learn positive/negative classification models from the training data. We expect that the main challenge will be creating good feature vectors to represent each example. As a starting point in the search for resources, we recommend the *ACL Resources List* [11].

References

[1] Relation 7: Training Data, http://docs.google.com/View?docID=w.df735kg3_8gt4b4c

[2] SemEval 2007: 4th International Workshop on Semantic Evaluations, http://nlp.cs.swarthmore.edu/semeval/

[3] Classification of Semantic Relations between Nominals: Description of Task 4 in SemEval 2007, http://docs.google.com/View?docID=w.d2jm3f3_98kcwd4

[4] Google Groups: Semantic Relations, http://groups.google.com/group/semanticrelations

[5] SemEval-2007: Schedule, http://nlp.cs.swarthmore.edu/semeval/schedule.shtml

[6] Relation 7: Content-Container, http://docs.google.com/View?docID=w.df735kg3_3gnrv95

[7] WordNet Reference Manual: Format of Sense Index File, http://wordnet.princeton.edu/man/senseidx.5WN

[8] WordNet: A Lexical Database for the English Language, http://wordnet.princeton.edu/

[9] Relation 7: Queries, http://docs.google.com/View?docID=w.df735kg3_12dpk9mx

[10] Creative Commons Attribution-Share Alike 2.5 License, http://creativecommons.org/licenses/by-sa/2.5/

[11] ACL Resources List, http://aclweb.org/aclwiki/index.php?title=Resources