# Metonymy Resolution at SemEval I:<br>Guidelines for Participants

Katja Markert and Malvina Nissim
markert@comp.leeds.ac.uk, malvina.nissim@unibo.it

February 25, 2007

## Contents

# 1 Contact Details

This task is organised by Katja Markert, School of Computing, University of Leeds, UK (`markert@comp.leeds.ac.uk`) and Malvina Nissim, University of Bologna, Italy and Institute for Cognitive Science and Technology, CNR, Italy (`malvina.nissim@unibo.it`). Please contact the organisers as early as possible if you have any questions about the task, submissions or these guidelines.

Further details about the competition are available from both the task web-site (`http://www.comp.leeds.ac.uk/markert/MetoSemeval2007.html`) and the official Semeval site (`http://nlp.cs.swarthmore.edu/semeval`).

We also recommend that you sign up to the metosemeval07 newsgroup from the task web-site. *NEW*

# 2 Task Summary

Metonymy is a form of figurative speech, in which one expression is used to refer to the standard referent of a related one. For example, in Example (1), ”Vietnam”, the name of a location, refers to an event (a war) that happened there.

(1) at the time of **Vietnam**, increased spending led to inflation

Similar in the Examples (2) and (3), ”BMW” , the name of a company, stands for its index on the stock market or a vehicle manufactured by BMW, respectively.

(2) **BMW** slipped 4p to 31p.

(3) The **BMW** slowed down.

The importance of resolving (i.e. recognising and interpreting) metonymies has been shown for a variety of NLP tasks, such as machine translation (**?**), question answering (**?**), and anaphora resolution (**??**).

Although metonymic readings are potentially open-ended and can be innovative, most of them tend to follow specific patterns. Many other location names, for instance, can be used in the same fashion as “Vietnam” in Example (1) above. Thus, given a semantic class (such as location), one can specify several regular metonymic patterns (such as place-for-event) that instances of the class are likely to undergo.

The task is a lexical sample task for English, concentrating on the two semantic classes *location*, exemplified by country names, and *organisation*, exemplified by company names. Participants have to automatically classify preselected country/company names as having a literal or non-literal meaning, given a four-sentence context. Apart from this basic coarse-grained recognition of non-literal readings, participants can choose finer-grained interpretation levels, further classifying readings into prespecified metonymic patterns (such as place-for-event or company-for-stock) or as innovative metonymies. The following Section describes all target categories (including all metonymic patterns) for both semantic classes whereas Section 5 explains the different levels of interpretation granularity participants can choose.

# 3 Data Collection and Annotation

The training sets consist of a collection of text snippets from the British National Corpus (BNC), Version 1.0, coded in XML (see `http::/www.natcorp.ox.ac.uk` for more info about the BNC). Each snippet is called a *sample* and contains a country or company name annotated for metonymy. The complete training sets containing 925 samples for country names (`countries.train`), and 1090 samples for company names (`companies.train`) are included at the top level of the trial/training distribution. Apart from the test files, these are the **only** files that are definitely necessary to participate in this challenge.

Test sets have the same structure, format and source. The complete testing sets contain 908 samples for country names (`countries.test`), and 842 samples for company names (`companies.test`) and are included at the top level of the test distribution. *NEW*

## 3.1 Data Collection and Annotation Process

We used the CIA Factbook and the Fortune 500 list as sampling frames for country and company names respectively. All occurrences of all names in the sampling frames were extracted in context from the BNC, a 100 million word, balanced corpus of contemporary English. We used the tool Gsearch (?) for extraction, allowing for both upper and lower case occurrences as well as the occurrence of plural forms. Any text in the BNC was allowed to occur.

All samples contain up to four sentences: two sentences of context before the sentence in which the country/company name occurs, the sentence in which the country/company name occurs and one sentence after. In case the name occurs at the beginning or end of a text, the samples may contain less than four sentences.

A number of these samples was then randomly selected as training sets and manually annotated for metonymy. Each sample was annotated independently by both organisers.

First, several samples were removed from the dataset due to the following reasons:

- In rare cases, the sample was not understood by the annotators because of insufficient context
- The name occurrence extracted was a homonym not in the desired semantic class (for example, *Mr. Greenland* when looking for locations). This task is not about standard Named Entity Recognition, so we assume that the general semantic class of the name is already known.

For all those occurrences that do have the semantic class `location` or `org`, metonymy annotation was carried out, using the categories described in the following Subsection. Annotation was strictly checked for validity and reliability (?). Samples whose readings could not be agreed on (after a reconciliation phase) were excluded from the dataset. No attempt was made to stratify the reading distribution in the datasets — the distribution mirrors the actual distribution in the corpus.

### 3.2 Annotation Categories

We provide here a brief description of the readings we used and that we ask automated systems to distinguish. *The following description should be sufficient for participation.*

Obviously the annotation scheme incorporates prior insights on metonymy by us and other researchers. If you are interested in further details on linguistic background, rationale, scheme development, and testing of the validity and reliability of the annotation scheme we refer you to **?**. A more detailed description of the annotation categories and the annotation procedure (together with additional examples, and decision making instructions for annotators), is also given in the `full_annotationguide` directory enclosed in the distribution.

**Categories for Locations**

`literal` The literal reading for location names comprises a *locative* (see Example (4)) and a *political* entity interpretation (see Example (5)).

(4) coral coast of **Papua New Guinea**.

(5) **Britain**'s current account deficit.

`metonymic` metonymic readings comprise patterns as well as a category for innovative metonymies:

`place-for-people` a place stands for any persons/organisations associated with it. These could be the government, as in (Example 6), some organisation like a sports team as in Example (7) or the whole population as in Example (8). The referent can also be left underspecified as in Example (9).

(6) **America** did once try to ban alcohol.

(7) **England** lost in the semi-final.

(8) The notion that the incarnation was to fulfil the promise to **Israel** and to reconcile the world with God.

(9) The G-24 group expressed readiness to provide **Albania** with food aid.

`place-for-event` a location name stands for an event that happened in the location (see Example (1)).

`place-for-product` a place stands for a product manufactured in the place,

(10) a smooth **Bordeaux** that was gutsy enough to cope with our food

`object-for-name` all names can be used as mere signifiers, instead of referring to an object or set of objects. In Example (11), "Guyana" would receive a literal interpretation, whereas "British Guiana" is a mere reference to a previous name of the location.

(11) Guyana (formerly **British Guiana**) gained independence.

`object-for-representation` A proper noun can refer to a representation (such as a photo, a painting or a symbol) of the referent of its literal reading. Thus "Malta" in Example (12) refers to a drawing of the island when pointing to a map.

(12) This is **Malta**

`othermet` a metonymy that does not fall into any of the prespecified patterns, as in Example (13), where “New Jersey” refers to typical local tunes.

(13) The thing about the record is the influences of the music. The bottom end is very New York/**New Jersey** and the top is very melodic.

`mixed` similar to zeugma, sometimes two predicates are involved, triggering a different reading each. In Example (14), both a `literal` reading (triggered by “in”) and a `place-for-people` reading (triggered by “leading critic”) are involved.

(14) they arrived in **Nigeria**, hitherto a leading critic of [. . . ]

### Categories for Organisations

`literal` The literal reading for organisation names describes references to the organisation as a legal entity in general. Examples of literal readings include (among others) descriptions of the structure of an organisation (see Example (15)), associations between organisations (see Example (16)) or relations between organisations and products/services they offer (see Example (17)).

(15) **NATO** countries

(16) **Sun** acquired that part of Eastman-Kodak Co’s Unix subsidiary

(17) **Intel**’s Indeo video compression hardware

`metonymic` metonymic readings again comprise metonymic patterns as well as a category for innovative readings.

`organisation-for-members` an organisation stands for its members. For example a spokesperson or official acts or speaks for the organisation, as in Example (18), or all members of the organisation participate in an action, as in Example (19).

(18) Last February **IBM** announced [. . . ]

(19) It’s customary to go to work in black or white suits. [. . . ] **Woolworths** wear them

`organisation-for-event` an organisation name is used to refer to an event associated with the organisation (e.g. a scandal or bankruptcy), as in Example (20).

(20) A remarkable example of back-bench influence on the Prime Minister was seen in the resignation of Leon Brittan from Trade and Industry in the aftermath of **Westland**.

`organisation-for-product` frequently the name of a commercial organisation is used to refer to its products, as in Example (21).

(21) A red light was hung on the **Ford**’s tail-gate.

`organisation-for-facility` organisations can also stand for the facility that houses the organisation or one of its branches, as in the following example.

(22) The opening of a **McDonald’s** is a major event

`organisation-for-index` an organisation name can be used for an index that indicates its value, for example its stock index, as in Example (3) above.

`object-for-name` In Example (23), both Chevrolet and Ford are used as strings, rather than referring to the companies.

(23) **Chevrolet** is feminine because of its sound (it's a longer word than **Ford**, has an open vowel at the end, connotes Frenchness).

`object-for-representation` In Example (24), "Shell" refers to the Shell logo.

(24) BT's pipes-of-Pan motif was, for him, somehow too British. Graphically, it lacked what King calls the world class of **Shell**.

`othermet` a metonymy that does not fall into any of the prespecified patterns, as in Example (25), where "Barclays Bank" stands for an account at the bank.

(25) funds [. . . ] had been paid into **Barclays Bank**.

`mixed` Similar to zeugma, two predicates are involved, triggering a different reading each. In Example (26), for instance, both an `organisation-for-index` ("slipping") and an `organisation-for-members` pattern ("confirming") are invoked.

(26) **Barclays** slipped 4p to 351p after confirming 3,000 more job losses.

## 4 Data Format

Figure 1 shows an example in the training file `companies.train`.

```
<sample id="samp8">
<bnc:title> Liberating communications </bnc:title>
<par>
The main competition to IBM came from a group of other US mainframe producers,
often referred to as the &bquo; seven dwarfs &equo; or the Bunch.
The B was for Burroughs and the H was for Honeywell &mdash; and such US companies
were slowly falling further behind IBM.
This became especially apparent when the US Department of Justice dropped its
anti-trust case against {\bf <annot><org reading="literal"> IBM </org></annot>} in
January 1982.
IBM was waiting for this decision and quickly became more aggressive at home and
abroad.
</par>
</sample>
```

Figure 1: Example sample from training file

The sample provides a sample-id number *samp8*. Afterwards the title of the BNC file that the sample was extracted from is given. The main text is presented within the <`par`> element, each sentence on a different line. The name to be classified is always in the <`annot`> element, inside another element which specifies the semantic class (<`org`> or <`location`>) with an attribute (`reading`) for the reading `literal`, `metonymic` or `mixed`. If the reading is metonymic, a further attribute (metotype) specifies the metonymic pattern or othermet for innovative metonymic readings. Additionally, a notes attribute can provide yet more detailed information about the annotation in free text, which is, however, not a target for classification.

Please note that there is **only one name per sample that needs to be classified**, though there might be more instances of say company names in the same sample. Only the name which is within the <annot> element must be assigned a target class.

Test files are provided in the same format but replace the gold standard reading value with the value "unknown".

## 5 Submission

Please follow the following submission guidelines closely. Apart from that please also follow the general participants guidelines at http://nlp.cs.swarthmore.edu/semeval/participants.shtml closely. We only differ in the time frame (see below). *NEW*

### Time Frame

*NEW*

Our time frame varies slightly from the SemEval general time frame. You can keep the trial/training data for as long as you want before submitting your results. As the task consists of two subtasks, you must submit your results **two weeks** after downloading the test data.

### Submission Types and Classification Levels

You can participate in the location or org task or both tasks. We encourage different methods to take part, for example supervised and unsupervised machine learning methods as well as linguistically oriented or knowledge representation frameworks, systems using the training data as well as systems using and/or creating other resources. We also allow partial submissions (for example, systems that would like to concentrate on names in particular grammatical functions only).

Teams can choose between three progressively more detailed levels of granularity for their system output, *coarse* (easiest), *medium*, or *fine* (most difficult). Classification categories depend on the chosen granularity level. The columns in the Tables 1 and 2 list all possible target categories for location and org classes, according to the three different levels of granularity. As an example, if you submit a medium system, only the output strings literal, metonymic or mixed are allowed for any sample.

For each task, SemEval organisers only allow **one submission** per team to avoid many similar systems by the same team taking part. Note that for this task one submission can contain up to 6 output files, due to 2 different subtasks and 3 different granularity levels. You are *not* allowed two different runs for the same subtask and granularity level. We are aware that the system details might differ (for example in feature selection) for the location or org subtask or for different granularity levels.

Table 1: Location categories and different levels of granularity

| *coarse* | *medium* | *fine* |
|---|---|---|
| literal | literal | literal |
| non-literal | mixed | mixed |
| | metonymic | othermet |
| | | object-for-name |
| | | object-for-representation |
| | | place-for-people |
| | | place-for-event |
| | | place-for-product |

Table 2: Organisation categories and different levels of granularity

| *coarse* | *medium* | *fine* |
|---|---|---|
| literal | literal | literal |
| non-literal | mixed | mixed |
| | metonymic | othermet |
| | | object-for-name |
| | | object-for-representation |
| | | organisation-for-members |
| | | organisation-for-event |
| | | organisation-for-product |
| | | organisation-for-facility |
| | | organisation-for-index |

## Submission Format

Please name your output files `systemidentifier.class.granularity`, where the system identifier corresponds to the name you gave on the SemEval registration site, the class name is either `location` or `org` and the granularity is *coarse*, *medium* or *fine*. Examples are *NLPuni.location.fine* or *NLPcompany.org.medium*.

Please provide each output file in the following format as a plain text file (no Word files). Sample submission files are included in the directory `SampleSubmissions`.

```
# <systemidentifier>
# <class> {location,org}
# <granularity> {coarse,medium,fine}
samp1|predictedreading
samp2|predictedreading
samp3|...
```

The possible values of the predicted reading depend on the level of granularity chosen (see above). You must provide exactly one output line for all samples in the test set. To allow for partial submissions, an additional value `unknown` can be used for cases that systems might not be able to cover, independent of granularity level.

Note that the sample number must be directly taken from each excerpt (e.g., <`sample id="samp8"`>). Please make sure that you do not renumber the samples. Apart from the first three lines, the system output can be provided in any order.

## 6 Evaluation

Systems will be evaluated according to standard measures. More specifically, they are divided into overall measures, and per-class measures.

**Overall measures** To assess the overall performance of a submitted system we calculate accuracy and coverage, allowing for partial submissions:

**accuracy** $\frac{\#\ correct\ predictions}{\#\ predictions}$

**coverage** $\frac{\#\ predictions}{\#\ samples}$

For partial submissions, we will also compute the accuracy which results if all `unknown` assignments are replaced by the baseline `literal`, yielding an output file with coverage 1. The resulting accuracy measures is called *allaccuracy*. For full submissions *allaccuracy* is equal to *accuracy*.

**Per-class measures** For each target category $c$, we calculate the following measures. The set of target categories depends on the chosen classification level.

**precision**$_c$ $\frac{\#\ correct\ assignments\ of\ c}{\#\ assignments\ of\ c}$

**recall**$_c$ $\frac{\#\ correct\ assignments\ of\ c}{\#\ dataset\ instances\ of\ c}$

**f-score**$_c$ $\frac{2 precision_c recall_c}{precision_c + recall_c}$

For partial submissions, we also compute the corresponding precision/recall and f-measure for an algorithm that replaces all `unknown` assignments with `literal`.

The evaluation script provided in the `Scripts` directory outputs all these measures, given a submission file (see Section 5) and a gold standard file in the format of `countries.train`. As it also checks a submission file for being well-formed, we encourage its usage.

We will compare the system output to a baseline that does not recognise any metonymies but just chooses `literal` for all samples. In our training sets that corresponds to the most frequent category (at each granularity level). It achieves an accuracy of 79.7% for `countries.train` and 63.3% for `companies.train`. The reading distribution in the test set might vary from the training set distribution.

In addition, we will also determine which readings are especially hard to distinguish and determine problem cases.

# A Appendix A: References

# B Appendix B: Statistics for Training Sets

The Tables 3 and 4 include the reading distributions in the training set data. As you can see, the companies data is more varied than the countries data. As some readings for locations are rare to find with the random sampling method, the subdirectory `ExtraSamples` contains a file `countries.extras` which contains some additional examples to illustrate some of the rarer readings.

Table 3: Distribution of readings in `companies.train`

| reading | N | % |
|---|---|---|
| literal | 690 | 63.3 |
| organisation-for-members | 220 | 20.2 |
| organisation-for-event | 2 | 0.2 |
| organisation-for-product | 74 | 6.8 |
| organisation-for-facility | 15 | 1.4 |
| organisation-for-index | 7 | 0.6 |
| object-for-name | 8 | 0.7 |
| object-for-representation | 1 | 0.1 |
| othermet | 14 | 1.3 |
| mixed | 59 | 5.4 |
| total | 1090 | 100.0 |

Table 4: Distribution of readings in `countries.train`

| reading | N | % |
|---|---|---|
| literal | 737 | 79.7 |
| place-for-people | 161 | 17.4 |
| place-for-event | 3 | 0.3 |
| place-for-product | 0 | 0.0 |
| object-for-name | 0 | 0.0 |
| object-for-representation | 0 | 0.0 |
| othermet | 9 | 1.0 |
| mixed | 15 | 1.6 |
| total | 925 | 100.0 |

# C Appendix C: SemEval Material and Directory Structure

The training/testing zip archives are organised in the following way. At the top level you will find all the files you need for the task. Subdirectories contain extra information which you might optionally want to use for developing your system.

The training/testing archives are structured in a parallel way. The following description separate training/testing structure by slashes. *NEW*

## C.1 Necessary Data Files

These are the only strictly necessary data files you need to take part in the challenge.

- `companies.train/companies.test`

  XML training/testing file for company names. The training file contains a total of 1090 samples; the testing file a total of 842 samples.

- `countries.train/countries.test`

  XML training/testing file for country names. The training file contains a total of 925 samples; the testing file a total of 908 samples.

- `SemEval.dtd/SemEvaltest.dtd`

  dtd for the above xml files.

Data format of these training/testing files is described in Section 4.

## C.2 Readme and Documentation Files

These are the documents you should definitely read before taking part in the challenge:

- `guidelines.pdf/guidelines.updated.pdf`

  this file

- `COPYRIGHT`

  copyright statement concerning data provenance and usage

## C.3 Optional Material

All other files contain additional material, providing different formats of the data, scripts or additional annotation that you might want to use for system development. *It is not necessary to read or use these files to take part in the competition.*

- `BNCtagged`

This directory contains the country and company datasets including the original BNC annotation (document meta-information, sentence and word boundaries and POS tags for all words). The metonymy annotation is knitted in the BNC XML files. There are four files:

  - `companies.train.BNCtagged/companies.test.BNCtagged` XML training/testing file for the company names dataset with BNC and metonymy annotation
  - `countries.train.BNCtagged/countries.test.BNCtagged` XML training/testing file for the country names dataset with BNC and metonymy annotation
  - `SemEval.BNCtagged.dtd/SemEvaltest.BNCtagged.dtd` dtd for XML files above
  - `README`: a short README describing the format

- `BNCtagged-standoff`

  This directory contains the country and company names datasets in the following format: XML files for the BNC texts including the original BNC annotation (metainformation, sentence/word boundaries, and POS tags), and separate files with the metonymy annotation as standoff pointing to the XML BNC files. There are six files, two per dataset plus a dtd and a README:

  - `companies.train.BNCtagged.samples/companies.test.BNCtagged.samples` XML BNC text file for the company names dataset
  - `companies.train.BNCtagged.annotation/companies.test.BNCtagged.annotation` XML standoff metonymy annotation file for the company names dataset (points to XML BNC file above)
  - `countries.train.BNCtagged.samples/countries.test.BNCtagged.samples` XML BNC text file for the country names dataset
  - `countries.train.BNCtagged.annotation/countries.test.BNCtagged.annotation` XML standoff metonymy annotation file for the country names dataset
  - `SemEval.BNCtagged.dtd/SemEvaltest.BNCtagged.dtd` dtd for all XML files above.
  - `README` a short README describing the format

- `grammannot`

  As optional information that teams might want to use, we include in this directory the grammatical relations that the name to be classified is involved in. The relations were annotated manually, using a dependency framework. Four files are included in this directory:

  - `companies.train.grammannot/companies.test.grammannot` grammar manual annotation for the company dataset
  - `countries.train.grammannot/countries.test.grammannot` grammar manual annotation for the country dataset
  - `README` file explaining format of grammar annotation.

  - `scheme.grammar.pdf`: Detailed guidelines on the dependency relations used.

- `ExtraSamples`

  As some readings for locations are rare to find with the random sampling method, the subdirectory `ExtraSamples` in the training archive contains a file `countries.extras` which contains some additional examples to illustrate some of the rarer readings.

- `Scripts`

  `remove-characters.pl` A perl script that can be used to replace BNC character entities in the samples. Thus, for example, "&mdash;" will be replaced by "—".

  `eval.pl` Perl script which takes as input two files (system output and gold standard), and evaluates the system according to the evaluation measures in Section 6.

- `SampleSubmissions`

  One or more sample submission files exemplifying the submission format.

- `full_annotationguide`

  This directory contains more detailed guidelines of the annotation categories, as used by the annotators during annotation.