UP13: Knowledge-poor Methods (Sometimes) Perform Poorly

Thierry Poibeau
Laboratoire d'Informatique de Paris-Nord
CNRS UMR 7030 et université Paris 13
99, avenue J.-B. Clément F-93430 Villetaneuse
thierry.poibeau@lipn.univ-paris13.fr

Abstract

This short paper presents a system developed at the Université Paris 13 for the Semeval 2007 Metonymy Resolution Task (task #08, location name track; see Markert and Nissim, 2007). The system makes use of plain word forms only. In this paper, we evaluate the accuracy of this minimalist approach, compare it to a more complex one which uses both syntactic and semantic features, and discuss its usefulness for metonymy resolution in general.

1 Introduction

This short paper presents the system developed at the Université Paris 13 for the Metonymy resolution task, during Semeval 2007 (Markert and Nissim, 2007). Two sub-tasks were proposed, concerning 1) country names and 2) company names. We only participated in the first task (country names). We developed a simple approach which we present and thoroughly evaluate in this paper. We discuss the relevance of this approach and compare it to more complex ones.

2 Motivation

We participated in the metonymy task with a very basic system. The idea was to investigate the efficiency of a minimalist (though, not Chomskian) system. This system tags entities on the basis of discriminative (plain) word forms occurring in a given window only. Our aim was to find out which word forms are discriminative enough to be considered as parameters.

In the past, we developed a system for metonymy resolution for French, evaluated in the framework of the ESTER evaluation (Gravier, 2004). This system, described in Poibeau (2006), uses various kinds of information, among others: plain word forms, part-of-speech tags, and syntactic and semantic tags (conceptual word classes).

The usefulness of complex linguistic features (especially syntactic and semantic tags) is questionable: they may be hard to compute, error-prone and their contribution is not clear. We therefore developed a new version of the system mainly based on 1) a distributional analysis (on surface word forms) along with 2) a filtering process. The latter restricted metonymic readings to country and capital names (as opposed to other location names), since they include a vast majority of the metonymic readings (this proved to be efficient but is of course a harsh pragmatic over-simplification without real linguistic basis). We nevertheless obtained a highly versatile system, performing reasonably well, compared to our previous, much more complex implementation (F-score was .58 instead of .63; we computed F-score with $\beta=1$).

In the framework of the Semeval evaluation, the filtering process is irrelevant since only country names are considered as entities. However, we thought that it would be interesting to develop a very basic system, to evaluate the performance one can obtain using plain word forms only.

3 A (too) Lazy Approach

We chose not to use any part-of-speech tagger or syntactic or semantic analyzer; we did not use any external knowledge or any other annotated corpus than the one provided for the training phase. Since no NLP tool was used, we had to duplicate most of the words in order to get the singular and the plural form. Our system is thus very simple compared to

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 418–421,
Prague, June 2007.

the state-of-art in this domain (e.g. Nissim and Markert, 2003).

We used discriminative plain words only. These are gathered as follows: all the words in a given window (here we use a 7 word window, before and after the target entity since it gave the best results on the training data) are extracted and associated with two classes (literal *vs.* non literal). We thus consider the most discriminative words, *i.e.* words that appear frequently in some contexts but not in others (literal *vs.* non-literal readings). Discriminative words are elements that are abnormally frequent or rare in one corpus compared to another one.

Characteristic features are selected based on their probabilities. Probability levels measure the significance of the differences between the relative frequency of an expression or a feature within a group (or a category) with its global relative frequency calculated over the entire corpus (Lafon, 1980). They are calculated under the hypothesis of a random distribution of the forms. The smaller the probability levels, the more characteristic the corresponding forms (Lebart and Salem, 1997).

We thus obtained 4 lists of discriminative words (literal *vs.* non-literal × before *vs.* after the target entity). As the result, some semantic families emerged, especially for words appearing before literal readings: lists of prepositions (*in*, *at*, *within*…) and geographical items (*east*, *west*, *western*…). Some lists were manually completed, when a "natural" series appeared to be incomplete (for example, if we got *east*, *west*, *north*, we completed the word series with *south*).

3.1 Reducing the Size of the Search Space

The approach described so far may seems a bit simplistic (and, indeed, it is!), but nevertheless it yielded highly discriminative features. For example, if we only tag country names immediately preceded by the preposition *in* as 'literal', we obtain the results presented in table 1 (in the following tables, precision is the most relevant issue; coverage gives an idea of the percentage of tagged entities by the considered feature, compared to the total number of entities to be tagged). Figure 1 shows that detecting the preposition *in* in front of a location name discriminates almost perfectly 23% of the literal readings.

	Training	Test
Precision	1	.98
Coverage	.23	.23

Table 1. Results for the pattern `in + LOC` (result tag = literal)

A simple discriminative analysis of the training corpus produces the following list of prepositions and geographical discriminative features: `"at"`, `"within"`, `"in"`, `"into"`, `"from"`, `"coast"`, `"land"`, `"area"`, `"southern"`, `"south"`, `"east"`, `"north"`, `"west"`, `"western"`, `"eastern"`, etc[1]. Table 2 presents the results obtained from this list of words (occurring in a 7 word window, on the left of the target word):

	Training	Test
Precision	.91	.88
Coverage	.60	.55

Table 2. Results for the pattern `<at+within+…> + LOC` (note that table 1 is contained in table 2)

Another typical feature was the use of the entity in a genitive construction (e.g. in *Iran's official commitment*, *Iran* is considered as a literal reading). The presence of *'s* on the right side of the target entity is highly discriminative (table 3):

	Training	Test
Precision	.87	.89
Coverage	.15	.17

Table 3. Results for the pattern `LOC's` (result tag = literal)

This strategy may seem strange, since the task is to find metonymic readings rather than literal ones (the baseline is to tag all the target entities as literal). However, it is useful in reducing the size of the search space by approximately 50%. This means that more than 70% of the entities with a literal meaning can be tagged with a confidence around 90% using this technique, thus reducing the number of problematic cases. The resulting file is relatively balanced: it contains about 50-60% of literal meaning and 40-50% of metaphorical meaning (instead of a classical ratio 80% *vs.* 20%).

[1] The list also contains nouns and verbs like: `"enter"`, `"entered"`, `"fly"`, `"flown"`, `"went"`, `"go"`, `"come"`, `"land"`, `"country"`, `"mountain"`…

3.2 Looking for Metonymy, Desperately …

We used the same strategy for metonymic readings. We have observed in the past that word forms are much more efficient for literal readings than for metonymic readings. However, the fact that the location name is followed by a verb like "has", "should", "was", "would", "will" seemed to be discriminative on the training corpus. Unfortunately, this feature did not work well on the test corpus (table 4).

	Training	Test
Precision	.6	.3
Coverage	.1	.04

Table 4. Results for the pattern LOC + <was, should…> (result tag = metonymic)

This simply means that a syntactic analysis would be useful to discriminate between the sentences where the target entity is the subject of the following verb (in this context, the entity is most of the time used with a metaphoric reading; to go further, one needs to filter the verb according to semantic classes).

Another point that was clear from the task guidelines was that sport's teams correspond to metonymic readings. The list of characteristic words for this class, obtained from the training corpus was the following: "player", "team", "defender", "plays", "role", "score", "scores", "scored", "win", "won", "cup", "v"[2], "against", "penalty", "goal", "goals", "champion", "champions", *etc*. But, bad luck! This list did not work well on the test corpus either:

	Training	Test
Precision	.64	.32
Coverage	.13	.05

Table 5. Results for the pattern LOC + <player, team…> (result tag = metonymic)

Table 5 shows that coverage as well as precision are very low.

Yet another category included words related to the political role of countries, which entails a metonymic reading: "role", "institution", "preoccupation", "attitude", "ally", "allies", "institutions", "initiative", "according", "authority"… All these categories had low coverage on the test corpus. This is not so surprising and is related to our knowledge-poor strategy: the training corpus is relatively small and it was foreseeable that we would miss most of the relevant contexts. However, we wanted to maintain precision above .5 (*i.e.* relevant contexts should remain relevant), but failed in this, as one can see from the overall results.

4 Overall Evaluation

We mainly discuss here the results of the *coarse* evaluation, where only literal *vs* non-literal meanings were targeted. We did not develop any specific strategy for the other tracks (*medium* and *fine*) since there were too few examples in the training data. We just transferred non-literal readings to the most probable class according to the training corpus (metonymic for *medium*, place-for-people for *fine*). However, the performance of our system (*i.e.* accuracy) is relatively stable between these three tracks, since the distribution of examples between the different classes is very unequally distributed.

Before giving the results, recall that our purpose was to investigate a knowledge-poor strategy, in order to establish how far one can get using only surface indicators. Thus, unsurprisingly, our results for the training corpus were already lower than those obtained using a more sophisticated system (Nissim and Markert, 2003). They are however a good indicator of performance when one uses only surface features.

The accuracy on the training corpus was .815. Precision and recall are presented in the table 6.

	Literal	Non-lit.
Precision	.88	.54
Recall	.88	.57
P&R	.88	.55

Table 6. Overall results on the training corpus

Accuracy on the test corpus is .754 only. Table 7 shows the results obtained for the different kinds of location names. The result is obvious: there is a significant drop in both recall and precision, compared to the results on the training corpus.

[2] *v* for *versus*, especially in sports: *Arsenal-MU 3 v 2*.

	Literal	Non lit.
Precision	.83	.38
Recall	.86	.31
P&R	.84	.34

Table 7. Overall results on the test corpus

5 Discussion

Metonymy is a complex linguistic phenomenon and it is thus no surprise that such a basic system performed badly, even if the drop in precision between training and test set was disappointing. The main conclusion of this approach is that surface forms can be used to reduce the size of the search space with a relatively good accuracy. A large part of the literal readings can be tagged using surface forms only. For the remaining cases, the use of more sophisticated linguistic information (both syntactic and semantic) is necessary.

During this work, we discovered some problematic target entities whose annotation is challenging. For instance, we tagged the following example as metonymic (because of the keywords "role" and "above"), whereas it is tagged as literal in the gold standard:

```
This two-track approach was seen (…) as
reflecting continued manoeuvring over
the role of the <annot> <location
reading="literal"> United States
</location> </annot> in the alliance, …
```

See also the following example (tagged by our system as metonymic because of the keyword "relations", but assumed to be literal in the gold standard):

```
Relations with China and <annot>
<location reading="literal"> Singapore
</location></annot> …
```

On the other hand, the following example was tagged as literal by our system (due to the preposition *in*) instead of metonymic.

```
After their European Championship
victory (…), Holland will be expected
to do well in <annot> <location
reading="metonymic" metotype="place-
for-event"> Italy </location></annot>.
```

If *Italy* is assumed to refer to the World Cup occurring in Italy, we think that the literal reading is not completely irrelevant (a paraphrase could be: *"…to do well during their stay in Italy"* which is clearly literal).

Metonymy is a form of figurative speech "in which one expression is used to refer to the referent of a related one" (Markert and Nissim, 2007). The phenomenon corresponds to a semantic shift in interpretation ("a profile shift") that appears to be a function of salience (Cruse and Croft, 2004). We assume that this semantic shift does not completely erase the original referent: it rather puts the focus on a specific feature of the content ("the profile") of the standard referent. If we adopt this theory, we can explain why it may be difficult to tag some examples, since both readings may co-exist.

6 Conclusion

In this paper, we presented a (minimalist) system for metonymy resolution and evaluated its usefulness for the task. The system worked well for reducing the size of the search space but performed badly for the recognition of metonymic readings themselves. It should be used in combination with more complex features, especially syntactic and semantic information.

References

A. Cruse and W. Croft. 2004. *Meaning in language, an introduction to semantics and pragmatics*. Oxford University Press, Oxford.

G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait and K. Choukri. 2004. The ESTER evaluation campaign for the rich transcription of French broadcast news". *Proceedings of LREC'04*. Lisbon, Portugal. pp. 885–888.

P. Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*. 1. pp. 127–165.

L. Lebart and A. Salem. 1997**.** *Exploring Textual Data*. Springer. Berlin.

K. Markert and M. Nissim. 2007. Task08: Metonymy Resolution at Semeval 2007. *Proceedings of Semeval 2007*. Prague, Czech Rep.

M. Nissim and K. Markert. 2003. Syntactic Features and Word Similarity for supervised Metonymy Resolution. *Proceedings of ACL'03*. Sapporo, Japan. pp. 56–63.

T. Poibeau. 2006. Dealing with Metonymic Readings of Named Entities. *Proceedings of COGSCI'06*. Vancouver, Canada. pp. 1962–1968.