UP13: Knowledge poor methods (sometimes) perform poorly

Thierry Poibeau

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 Viletaneuse
thierry.poibeau@lipn.univ-paris13.fr

Abstract

This short paper presents the system developed at the Université Paris 13 for the Metonymy resolution task, during Semeval 2007 (location name track). We developed a basic strategy only based on plain word forms to see how far one can go using only surface cues. We then discuss the relevance of this approach and compare it with more complex ones.

1 Introduction

This short paper presents the system developed at the Université Paris 13 for the Metonymy resolution task, during Semeval 2007. Two sub-tasks were proposed, concerning 1) country names and 2) location names: we only participated to the first track (country names). We developed a basic strategy that is presented and thoroughly evaluated. We then discuss the relevance of this approach and compare it with more complex ones.

2 Motivation

We participated to the metonymy task with a very basic system, developed in one day (but maybe some more days would not have been superfluous…). The idea was to see how far one can go with a minimalist (through, not Chomskian) system. The principle was to tag entities on the basis of discriminative (plain) word forms occurring in a given window. Our aim was then to discover which word forms were discriminative enough to be considered as relevant parameters.

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

The need for complex linguistic features (especially syntactic and semantic tags) is problematic: they may be hard to compute, error-prone and their contribution is not clear. We then realized a new version of the software mainly based on 1) a distributional analysis (on surface word forms) along with 2) a filtering process (only country or capital names can have a metonymic reading, as opposed to other location names). Using these (over-simplified) features, we obtain a highly versatile system, performing not so badly compared to our previous, much more complex, implementation (.58 P&R instead of .63; P&R is the harmonic mean of precision and recall).

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

3 A (too) lazy approach

We did not use any part-of-speech tagger, nor any syntactic or semantic analyzer; we did not use any external knowledge nor any other annotated corpus than the one provided for the training phase. Since we decided not to use any NLP tool, 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 the state-of-art in this domain (e.g. Nissim and Markert, 2003).

We only used discriminative plain words. These are computed as follows: all the words in a given window (here we use a 8 word window, before and

after the target entity) are extracted and classified in two classes (literal *vs* non literal). We thus compute the most discriminative words, wrt. words that appear frequently in a context but not in the other (literal *vs* non-literal),. Discriminative words are elements that are abnormally frequent or abnormally rare in a corpus compared to another one.

Probability levels are used to select these characteristic features. The 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 computed on the whole corpus (Lafon, 1980). They are computed under the hypothesis of a random distribution of the form under consideration in the categories. The smaller are the probability levels, the more characteristic are the corresponding forms (Lebart and Salem, 1997).

We thus obtain 4 lists of discriminative words (literal *vs* non-literal × before *vs* after the target entity). Some semantic families automatically emerged from the analysis, especially among 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 get *east*, *west*, *north*, we completed the word series with *south*).

3.1 Reducing the size of the search space

This approach described so far may seem a bit too simplistic (and, indeed, it is!), but nevertheless we observed very discriminative features. For example, if we only tag country names immediately preceded by the preposition *in* as 'literal', we obtain the following results (in these 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):

	Training	Test
Precision	1	.98
Coverage	.23	.23

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

In other words, detecting the preposition *in* in front of a location's name discriminate quite perfectly 23% of the literal readings.

From the training corpus, a simple discriminative analysis provide 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]. From this list of words (occurring in a 8 word window, on the left of the target word), we obtain the following results:

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

Tab 2. Results for the pattern `<at+within+…> + LOC` (note that tab1 is a subpart of tab2)

Another typical feature was the use of the entity in a genitive construction (e.g. `<annot><location reading="literal"> Iran </location> </annot> 's official commitment`). The presence of `'s` on the right side of the target entity is also highly discriminative:

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

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

This strategy may seem misleading, since the task consists in finding metonymic readings rather than literal ones (the baseline consists in tagging everything as literal). However, it allows reducing the size of the search space by approximately 50% (i.e. more than 70% of the entities with a literal meaning can be tagged with a good confidence 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 verbs and nouns like: `"enter"`, `"entered"`, `"fly"`, `"flown"`, `"went"`, `"go"`, `"come"`, `"land"`, `"country"`, `"countries"`, `"northern"`, `"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's name is followed by a verb like "has", "should", "was", "would", "will" seemed to be discriminative on the training corpus.

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

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

Unfortunately, this feature did not work well on the test corpus. 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", "against", "penalty", "goal", "goals", "champion", "champions"…` But, bad luck! This list did not work well on the test corpus either:

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

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

Coverage as well as precision is 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 a low coverage on the test corpus. This is not so surprising and is related to our poor strategy: the training corpus is relatively small and it was foreseeable that we would miss most of the relevant contexts. However, we were optimistically planning to maintain precision above .5 (*i.e.* relevant contexts should remain relevant), which was not the case, as one can see from the overall results.

4 Overall evaluation[2]

Before giving the overall results, let's remind the reader that we wanted to test a knowledge poor strategy, to check how far we can go using only surface indicators. Thus, even the results obtained from the training corpus were not comparable to what is obtained from more complex knowledge source (Nissim and Markert, 2003).

Accuracy on the training corpus was .815. Precision and recall are presented in the following table.

	Lit	Met
Precision	.88	.54
Recall	.88	.57
P&R	.88	.55

Tab 6. Overall results on the training corpus

Accuracy on the test corpus is .754 only. The following results were obtained for the different kinds of location's names:

	Lit	Met
Precision	.83	.38
Recall	.86	.31
P&R	.84	.34

Tab 7. Overall results on the test corpus

The result is obvious: there is an important drop both in recall and in precision, compared to performances obtained on the training corpus.

[2] We only discuss here the results obtained for the *coarse* evaluation, where only `literal` *vs* `non-literal` meaning had to be found. We did not develop any specific rule for the other tracks (*medium* and *fine*) since there were too few examples. We just transfer non-literal readings on the most probable class (`metonymic` for *medium*, `place-for people` for *fine*). However, accuracy of our system is relatively stable between these three tracks, since the distribution of examples between these different classes is very unequally distributed.

5 Discussion

Metonymy is a complex linguistic phenomenon and it is no so surprising that a so basic system performed badly, even if we were disappointed by the drop of precision between training and test. The main conclusion of this approach is that, even if surface forms are acceptable to reduce the size of the search space with a relatively good accuracy, there are a large number of remaining cases for which other linguistic information (both syntactic and semantic) is necessary.

Note however that some examples are difficult and should be further discussed. 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, against a background of US troop reductions in Europe and Franco-German proposals for a European military force.

See also the following example (tagged by our system as metonymic because of the keyword "relations", but assumed to be literal from 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 and Milan's orange-tinted European Cup triumph, 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, I think that the literal reading is not completely irrelevant either (a paraphrase could be: *"…to do well during their stay in Italy"* which is clearly literal).

Metonymy is defined by the organizers as "a form of figurative speech, in which one expression is used to refer to the standard referent of a related one" (Markert and Nissim, 2007). This phenomenon corresponds to a semantic shift in interpretation ("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 put the focus on a specific feature of the profile of the standard referent. If we believe in this explanation, it explains why it is sometimes hard to decide how to tag some examples, since both readings may co-exist.

6 Conclusion

In this paper, we presented a basic (minimalist) system for metonymy resolution. The strategy worked well for reducing the size of the search space but performed badly for the recognition of metonymic readings themselves. If one still think that this strategy has some interest, it must at least be used combined with more complex features, especially syntactic and semantic information.

This conclusion is for sure not impressive but we hope to have given an idea of what is a kind of bottom line for the task since simple heuristics may work relatively well in other contexts (see the experiment on ESTER, section 2.2).

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". *Procedings of Language and Resource Evluation Conference* (LREC). Lisboa, 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. Metonymy Resolution at SemEval I: Guidelines for Participants.

M. Nissim and K. Markert. 2003. Syntactic Features and Word Similarity for supervised Metonymy Resolution. *Proceedings of Association for Computational Linguisitcs Conference* (ACL). Sapporo, Japan. pp. 56–63.

T. Poibeau. 2006. Dealing with Metonymic Readings of Named Entities. *Proceedings of Cognitive SScience* (COGSCI). Vancouver, Canada. pp. 1962–1968.