XRCE-M: A Hybrid System for Named Entity Metonymy Resolution

***Caroline Brun** ***Maud Ehrmann** ***Guillaume Jacquet**

* Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan France
*{Caroline.Brun, Maud.Ehrmann, Guillaume.Jacquet}@xrce.xerox.com

Abstract

This paper describes our participation to the Metonymy resolution at SemEval 2007 (task #8). In order to perform named entity metonymy resolution, we developed a hybrid system based on a robust parser that extracts deep syntactic relations combined with a non-supervised distributional approach, also relying on the relations extracted by the parser.

1 Description of our System

SemEval 2007 introduces a task aiming at resolving metonymy for named entities, for location and organization names (Markert and Nissim 2007). Our system addresses this task by combining a symbolic approach based on robust deep parsing and lexical semantic information, with a distributional method using syntactic context similarities calculated on large corpora. Our system is completely unsupervised, as opposed to state-of-the-art systems (see (Market and Nissim, 2005)).

1.1 Robust and Deep Parsing Using XIP

We use the Xerox Incremental Parser (XIP, (Aït et al., 2002)) to perform robust and deep syntactic analysis. Deep syntactic analysis consists here in the construction of a set of syntactic relations[1] from an input text. These relations, labeled with deep syntactic functions, link lexical units of the input text and/or more complex syntactic domains that are constructed during the processing (mainly chunks, see (Abney, 1991)). Moreover, together with surface syntactic relations, the parser calculates more sophisticated relations using derivational morphologic properties, deep syntactic properties[2], and some limited lexical semantic coding (Levin's verb class alternations, see (Levin, 1993)), and some elements of the Framenet[3] classification, (Ruppenhofer et al., 2006)). These deep syntactic relations correspond roughly to the agent-experiencer roles that is subsumed by the SUBJ-N relation and to the patient-theme role subsumed by the OBJ-N relation, see (Brun and Hagège, 2003). Not only verbs bear these relations but also deverbal nouns with their corresponding arguments.

Here is an example of an output (chunks and deep syntactic relations):

Lebanon still wanted to see the implementation of a UN resolution

```
TOP{SC{NP{Lebanon} FV{still wanted}} IV{to see} NP{the implementation} PP{of NP{a UN resolution}} .}
  MOD_PRE(wanted,still)
  MOD_PRE(resolution,UN)
  MOD_POST(implementation,resolution)
  COUNTRY(Lebanon)
  ORGANISATION(UN)
  EXPERIENCER_PRE(wanted,Lebanon)
  EXPERIENCER(see,Lebanon)
  CONTENT(see,implementation)
  EMBED_INFINIT(see,wanted)
  OBJ-N(implement,resolution)
```

1.2 Adaptation to the Task

Our parser includes a module for “standard” named entity recognition, but needs to be adapted to handle named entity metonymy. Following the guidelines of the SemEval task #8, we performed a

[1] inspired from dependency grammars, see (Mel’čuk, 1998), and (Tesnière, 1959).

[2] Subject and object of infinitives in the context of control verbs.

[3] http://framenet.icsi.berkeley.edu/

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 488–491,
Prague, June 2007. ©2007 Association for Computational Linguistics

corpus study on the trial data in order to detect lexical and syntactic regularities triggering a metonymy, for both location names and organization names. For example, we examined the subject relation between organizations or locations and verbs and we then classify the verbs accordingly: we draw hypothesis like "if a location name is the subject of a verb referring to an economic action, like *import*, *provide*, *refund*, *repay*, etc., then it is a place-for-people". We adapted our parser by adding dedicated lexicons that encode the information collected from the corpus and develop rules modifying the interpretation of the entity, for example:

```
If (LOCATION(#1) & SUBJ-N(#2[v_econ],#1))[4]
        ➔ PLACE-FOR-PEOPLE(#1)
```

We focus our study on relations like subject, object, experiencer, content, modifiers (nominal and prepositional) and attributes. We also capitalize on the already-encoded lexical information attached to verbs by the parser, like communication verbs like *say*, *deny*, *comment*, or categories of the FrameNet Experiencer subject frame, i.e. verbs like *feel*, *sense*, *see*. This information was very useful since experiencers denote persons, therefore all organizations or locations having an experiencer role can be considered as organization-for-members or place-for-people. Here is an example of output[5], when applying the modified parser on the following sentence:

*"It was the largest **Fiat** everyone had ever seen".*

```
  ORG-FOR-PRODUCT(Fiat)
  MOD_PRE(seen,ever)
  SUBJ-N_PRE(was,It)
  EXPERIENCER_PRE(seen,everyone)
  SUBJATTR(It,Fiat)
 QUALIF(Fiat,largest)
```

Here, the relation QUALIF(Fiat, largest) triggers the metonymical interpretation of "Fiat" as org-for-product.

This first development step is the starting point of our methodology, which is completed by a non-supervised distributional approach described in the next section.

1.3 Hybridizing with a Distributional Approach

The distributional approach proposes to establish a distance between words depending on there syntactic distribution.

The distributional hypothesis is that words that appear in similar contexts are semantically similar (Harris, 1951): the more two words have the same distribution, i.e. are found in the same syntactic contexts, the more they are semantically close.

We propose to apply this principle for metonymy resolution. Traditionally, the distributional approach groups words like *USA*, *Britain*, *France*, *Germany* because there are in the same syntactical contexts:

(1) Someone live in Germany.
(2) Someone works in Germany.
(3) Germany declares something.
(4) Germany signs something.

The metonymy resolution task implies to distinguish the literal cases, (1) & (2), from the metonymic ones, (3) & (4). Our method establishes these distinctions using the syntactic context distribution. We group contexts occurring with the same words: the syntactic contexts *live in* and *work in* are occurring with *Germany*, *France*, *country*, *city*, *place*, when syntactic contexts *subject-of-declare* and subject-of-*sign* are occurring with *Germany*, *France*, *someone*, *government*, *president*.

For each Named Entity annotation, the hybrid method consists in using symbolic annotation if there is (§1.2), else using distributional annotation (§1.3) as presented below.

Method: We constructed a distributional space with the 100M-word BNC. We prepared the corpus by lemmatizing and then parsing with the same robust parser than for the symbolic approach (XIP, see section 3.1). It allows us to identify triple instances. Each triple have the form w1.R.w2 where w1 and w2 are lexical units and R is a syntactic relation (Lin, 1998; Kilgarriff & *al.* 2004).

Our approach can be distinguished from classical distributional approach by different points.

First, we use triple occurrences to build a distributional space (one triple implies two contexts and two lexical units), but we use the transpose of the classical space: each point x_i of this space is a syntactical context (with the form R.w.), each dimension *j* is a lexical units, and each value $x_i(j)$ is the frequency of corresponding triple occurrences. Sec-

[4] Which read as "if the parser has detected a location name (#1), which is the subject of a verb (#2) bearing the feature "v-econ", then create a PLACE-FOR-PEOPLE unary predicate on #1.

[5] Only dependencies are shown.

ond, our lexical units are words but also complex nominal groups or verbal groups. Third, contexts can be simple contexts or composed contexts[6].

We illustrate these three points on the phrase *provide Albania with food aid*. The XIP parser gives the following triples where for example, *food aid* is considered as a lexical unit:

OBJ-N('VERB:provide','NOUN: Albania').
PREP_WITH('VERB: provide ','NOUN:aid').
PREP_WITH('VERB: provide ','NP:food aid').

From these triples, we create the following lexical units and contexts (in the context *1.VERB: provide. OBJ-N*, "*1*" mean that the verb *provide* is the governor of the relation OBJ-N):

Words:	Contexts:
VERB:provide	1.VERB: provide. OBJ-N
NOUN:Albania	1.VERB: provide.PREP_WITH
NOUN:aid	2.NOUN: Albania.OBJ-N
NP:food aid	2.NOUN: aid. PREP_WITH
	2.NP: food aid. PREP_WITH
	1.VERB:provide.OBJ-N+2.NOUN:aid. PREP_WITH
	1.VERB:provide.OBJ-N+2.NP:food aid. PREP_WITH
	1.VERB:provide.PREP_WITH +2.NO:Albania.OBJ-N

We use a heuristic to control the high productivity of these lexical units and contexts. Each lexical unit and each context should appear more than 100 times in the corpus. From the 100M-word BNC we obtained 60,849 lexical units and 140,634 contexts. Then, our distributional space has 140,634 units and 60,849 dimensions.

Using the global space to compute distances between each context is too consuming and would induce artificial ambiguity (Jacquet, Venant, 2005). If any named entity can be used in a metonymic reading, in a given corpus each named entity has not the same distribution of metonymic readings. The country *Vietnam* is more frequently used as an event than *France* or *Germany*, so, knowing that a context is employed with *Vietnam* allow to reduce the metonymic ambiguity.

For this, we construct a singular sub-space depending to the context and to the lexical unit (the ambiguous named entity):

For a given couple context *i* + lexical unit *j* we construct a subspace as follows:

Sub_contexts = list of contexts which are occurring with the word *i*. If there are more than k contexts, we take only the *k* more frequents.

Sub_dimension = list of lexical units which are occurring with at least one of the contexts from the Sub_contexts list. If there are more than *n* words, we take only the *n* more frequents (relative frequency) with the Sub_contexts list (for this application, *k* = 100 and *n* = 1,000).

We reduce dimensions of this sub-space to 10 dimensions with a PCA (Principal Components Analysis).

In this new reduced space (*k**10), we compute the closest context of the context *j* with the Euclidian distance.

At this point, we use the results of the symbolic approach described before as starting point. We attribute to each context of the Sub_contexts list, the annotation, if there is, attributed by symbolic rules. Each kind of annotation (literal, place-for-people, place-for-event, etc) is attributed a score corresponding to the sum of the scores obtained by each context annotated with this category. The score of a context *i* decreases in inverse proportion to its distance from the context *j*: score(context *i*) = 1/d(context *i*, context *j*) where d(*i*,*j*) is the Euclidian distance between *i* and *j*.

We illustrate this process with the sentence *provide Albania with food aid*. The unit *Albania* is found in 384 different contexts (|Sub_contexts| = 384) and 54,183 lexical units are occurring with at least one of the contexts from the Sub_contexts list (|Sub_dimension| = 54,183).

After reducing dimension with PCA, we obtain the context list below ordered by closeness with the given context (1.VERB:provide.OBJ-N):

Contexts	**d**	**symb. annot.**
1.VERB:provide.OBJ-N	0.00	
1.VERB:allow.OBJ-N	0.76	place-for-people
1.VERB:include.OBJ-N	0.96	
2.ADJ:new.MOD_PRE	1.02	
1.VERB:be.SUBJ-N	1.43	
1.VERB:supply.SUBJ-N_PRE	1.47	literal
1.VERB:become.SUBJ-N_PRE	1.64	
1.VERB:come.SUBJ-N_PRE	1.69	
1.VERB:support.SUBJ-N_PRE	1.70	place-for-people
etc.		

Score for each metonymic annotation of *Albania*:

→	**place-for-people**	**3.11**
	literal	1.23
	place-for-event	0.00
	…	0.00

The score obtained by each annotation type allows annotating this occurrence of *Albania* as a *place-for-people* metonymic reading. If we can't choose only one annotation (all score = 0 or equality between two annotations) we do not annotate.

[6] For our application, one context can be composed by two simple contexts.

2 Evaluation and Results

The following tables show the results on the **test** corpus:

type	Nb. samp	accuracy	coverage	Baseline accuracy	Baseline coverage
Loc/coarse	908	0.851	1	0.794	1
Loc/medium	908	0.848	1	0.794	1
Loc /fine	908	0.841	1	0.794	1
Org/coarse	842	0.732	1	0.618	1
Org/medium	842	0.711	1	0.618	1
Org/fine	842	0.700	1	0.618	1

Table 1: Global Results

	Nb occ.	Prec.	Recall	F-score
Literal	721	0.867	0.960	0.911
Place-for-people	141	0.651	0.490	0.559
Place-for-event	10	0.5	0.1	0.166
Place-for-product	1	_	0	0
Object-for-name	4	1	0.5	0.666
Object-for-representation	0	_	_	_
Othermet	11	_	0	0
mixed	20	_	0	0

Table 2: Detailed Results for Locations

	Nb occ.	Prec.	Recall	F-score
Literal	520	0.730	0.906	0.808
Organization-for-members	161	0.622	0.522	0.568
Organization-for-event	1	_	0	0
Organization-for-product	67	0.550	0.418	0.475
Organization-for-facility	16	0.5	0.125	0.2
Organization-for-index	3	_	0	0
Object-for-name	6	1	0.666	0.8
Othermet	8	_	0	0
Mixed	60	_	0	0

Table 3: Detailed Results for Organizations

The results obtained on the test corpora are above the baseline for both location and organization names and therefore are very encouraging for the method we developed. However, our results on the test corpora are below the ones we get on the train corpora, which indicates that there is room for improvement for our methodology.

Identified errors are of different nature:

Parsing errors: For example in the sentence "*Many galleries in the States, England and France declined the invitation.*", because the analysis of the coordination is not correct, *France* is calculated as subject of *declined*, a context triggering a place-for-people interpretation, which is wrong here.

Mixed cases: These phenomena, while relatively frequent in the corpora, are not properly treated.

Uncovered contexts: some of the syntactico-semantic contexts triggering a metonymy are not covered by the system at the moment.

3 Conclusion

This paper describes a system combining a symbolic and a non-supervised distributional approach, developed for resolving location and organization names metonymy. We plan to pursue this work in order to improve the system on the already-covered phenomenon as well as on different names entities.

References

Abney S. 1991. *Parsing by Chunks*. In Robert Berwick, Steven Abney and Carol Teny (eds.). Principle-based Parsing, Kluwer Academics Publishers.

Aït-Mokhtar S., Chanod, J.P., Roux, C. 2002. *Robustness beyond Shallowness: Incremental Dependency Parsing*. Special issue of NLE journal.

Brun, C., Hagège C., 2003. *Normalization and Paraphrasing Using Symbolic Methods*, Proceeding of the Second International Workshop on Paraphrasing. ACL 2003, Vol. 16, Sapporo, Japan.

Harris Z. 1951. *Structural Linguistics*, University of Chicago Press.

Jacquet G.,Venant F. 2003. *Construction automatique de classes de sélection distributionnelle,* In Proc. TALN 2003, Dourdan.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. 2004. *The sketch engine*. In Proc. EURALEX, pages 105-116.

Levin, B. 1993. *English Verb Classes and Alternations – A preliminary Investigation*. The University of Chicago Press.

Nissim, M. and Markert, K. 2005. *Learning to buy a Renault and to talk to a BMW: A supervised approach to conventional metonymy*. Proceedings of the 6th International Workshop on Computational Semantics, Tilburg.

Nissim, M. and Markert, K. 2007. *SemEval-2007 Task 08: Metonymy Resolution at SemEval-2007*. In Proceedings of SemEval-2007.

Lin D. 1998. *Automatic retrieval and clustering of similar words*. In COLING-ACL, pages 768-774.

Mel'čuk I. 1988. *Dependency Syntax*. State University of New York, Albany.

Ruppenhofer, J. Michael Ellsworth, Miriam R. L. Petruck, Christopher R Johnson and Jan Scheffczyk. 2006. *Framenet II: Extended Theory and Practice*.

Tesnière L. 1959. *Eléments de Syntaxe Structurale*. Klincksiek Eds. (Corrected edition Paris 1969).