An Application of WordNet to Prepositional Attachment

Sanda M. Harabagiu
University of Southern California
Department of Electrical Engineering-Systems
Los Angeles, CA 90089-2562
harabagi@usc.edu

Abstract

This paper presents a method for word sense disambiguation and coherence understanding of prepositional relations. The method relies on information provided by WordNet 1.5. We first classify prepositional attachments according to semantic equivalence of phrase heads and then apply inferential heuristics for understanding the validity of prepositional structures.

1 Problem description

In this paper, we address the problem of disambiguation and understanding prepositional attachment. The arguments of prepositional relations are automatically categorized into semantically equivalent classes of WordNet (Miller and Teibel, 1991) concepts. Then by applying inferential heuristics on each class, we establish semantic connections between arguments that explain the validity of that prepositional structure. The method uses information provided by WordNet, such as semantic relations and textual glosses.

We have collected prepositional relations from the Wall Street Journal tagged articles of the PENN TREEBANK. Here, we focus on preposition `of`, the most frequently used preposition in the corpus.

2 Classes of prepositional relations

Since most of the prepositional attachments obey the principle of locality (Wertmer, 1991), we considered only the case of prepositional phrases preceded by noun or verb phrases. We scanned the corpus and filtered the phrase heads to create $\mathcal{C}$, an ad hoc collection of sequences $<$ *noun prep noun* $>$ and $<$ *verb prep noun* $>$. This collection is divided into classes of prepositional relations, using the following definitions:

Definition 1: Two prepositional structures $<$ $noun_1$ *prep* $noun_2$ $>$ and $<$ $noun_3$ *prep* $noun_4$ $>$ belong to the same class if one of the following conditions holds:

- $noun_1$, and $noun_2$ are hypernym/hyponym of $noun_3$, and $noun_4$ respectively, or
- $noun_1$, and $noun_2$ have a common hypernym/hyponym and with $noun_3$, and $noun_4$, respectively.

A particular case is when $noun_1$ ($noun_2$) and $noun_3$ ($noun_4$) are synonyms.

Definition 2: Two prepositional structures $<$ $verb_1$ *prep* $noun_1$ $>$ and $<$ $verb_2$ *prep* $noun_2$ $>$ belong to the same class if one of the following conditions holds:

- $verb_1$, and $noun_1$ are hypernym/hyponym of $verb_2$, and $noun_2$, respectively or
- $verb_1$, and $noun_1$ have a common hypernym/hyponym with $verb_2$, and $noun_2$, respectively.

A particular case is when the verbs or the nouns are synonyms, respectively.

The main benefit and reason for grouping prepositional relations into classes is the possibility to disambiguate the words surrounding prepositions. When classes of prepositional structures are identified, two possibilities arise:

1. A class contains at least two prepositional sequences from the collection $\mathcal{C}$. In this case, all sequences in that class are disambiguated, because for each pair ($<$ $noun_i$ *prep* $noun_j$ $>$, $<$ $noun_k$ *prep* $noun_q$ $>$), $noun_i$ and $noun_k$ (and $noun_j$ and $noun_q$ respectively) are in one of the following relations:
   (a) they are synonyms, and point to one synset that is their meaning.
   (b) they belong to synsets that are in hypernym/hyponym relation.
   (c) they belong to synsets that have a common hypernym/hyponym.

   In cases (a), (b) and (c), since words are associated to synsets, their meanings are disambiguated. The same applies for classes of prepositional sequences $<$ *verb prep noun* $>$.

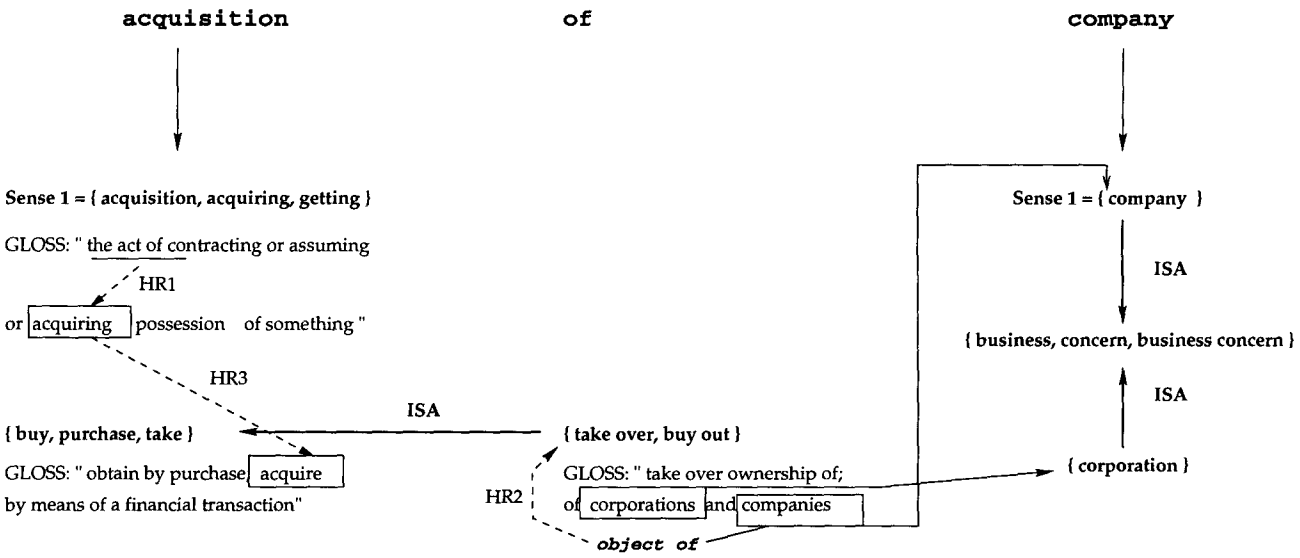

Figure 1: WordNet application of prepositional selection constraints

2. A class contains only one sequence. We disregard these classes from our study, since in this class it is not possible to disambiguate the words.

The collection $\mathcal{C}$ has 9511 $<$ *noun of noun* $>$ sequences, out of which 2158 have at least one of the nouns tagged as a proper noun. 602 of these sequences have both nouns tagged as proper nouns. Due to the fact that WordNet's coverage of proper nouns is rather sparse, only 34% of these sequences were disambiguated. Successful cases are $<$ *House of Representatives* $>$, $<$ *University of Pennsylvania* $>$ or $<$ *Museum of Art* $>$. Sequences that couldn't be disambiguated comprise $<$ *Aerospaciale of France* $>$ or $<$ *Kennedy of Massachusetts* $>$. A small disambiguation rate of 28% covers the rest of the 1566 sequences relating a proper noun to a common noun. A successful disambiguation occurred for $<$ *hundreds of Californians* $>$ or $<$ *corporation of Vancouver* $>$. Sequences like $<$ *aftermath of Iran-Contra* $>$ or $<$ *acquisition of Merryl Linch* $>$ weren't disambiguated. The results of the disambiguation of the rest of 7353 sequences comprising only common nouns are more encouraging. A total of 473 classes were devised, out of which 131 had only one element, yielding a disambiguation rate of 72.3%. The number of elements in a class varies from 2 to 68.

Now that we found disambiguated classes of prepositional structures, we provide some heuristics to better understand why the prepositional relations are valid. These heuristics are possible inferences performed on WordNet.

3 Selectional Heuristics on WordNet

In this section we focus on semantic connections between the words of prepositional structures. Consider for example `acquisition of company`. Figure 1 illustrates some of the relevant semantic connections that can be drawn from WordNet when analyzing this prepositional structure.

We note that noun `acquisition` is semantically connected to the verb `acquire`, which is related to the concept { `buy, purchase, take`}, a hypernym of { `take over, buy out`}. Typical objects for `buy out` are `corporations` and `companies`, both hypernyms of `concern`. Thus, at a more abstract level, we understand `acquisition of company` as an action performed on a typical object. Such relations hold for an entire class of prepositional structures.

What we want is to have a mechanism that extracts the essence of such semantic connections, and be able to provide the inference that the elements of this class are all sequences of $<$ *noun*$_i$ *prep noun*$_j$ $>$, with *noun*$_j$ always an object of the action described by *noun*$_i$.

Our approach to establish semantic paths is based on inferential heuristics on WordNet. Using several heuristics one can find common properties of a prepositional class. The classification procedure disambiguates both nouns as follows: the word `acquisition` has four senses in WordNet , but it is found in its synset number 1. The word `company` appears in its synset number 1. The gloss of `acquisition` satisfies the prerequisite of HR1:

Heuristic Rule 1 (HR1) If the textual gloss of a noun concept begins with the expression `the act of` followed by the gerund of a verb, then the respective noun concept describes an action represented by the verb from the gloss.

This heuristic applies 831 times in WordNet, showing that nouns like `accomplishment`, `dispatch` or `subsidization` describe actions.

Nr.crt.	Features for $< N1 >$ of $< N2 >$	Example
1	$N2$ is the object of the action described by $N1$	acquisition of company
2	$N2$ is the agent of the action described by $N1$	approval of authorities
3	$N1$ is the agent of the action with object $N2$	author of paper
4	$N1$ is the agent of the action with purpose the action described by $N2$	activists of support
5	$N1$ is the object of an action whose agent is $N2$	record of athlete
6	$N2$ describes the action with the theme $N1$	allegations of fraud
7	$N1$ is the location of the activity described by $N2$	place of business
8	$N1$ describes an action occurring at the time described by $N2$	acquisition of 1995
9	$N1$ is the consequence of a phenomenon described by $N2$	impact of earthquake
10	$N1$ is the output of an action described by $N2$	result of study

Table 1: Distribution of prepositions in the Wall Street Journal articles from PENN Treebank

Thus **acquisition** is a description of any of the verbal expressions **contract possession**, **assume possession** and **acquire possession**.

The role of **company** is recovered using another heuristic:

Heuristic Rule 2 (HR2) The gloss of a verb may contain multiple textual explanations for that concept, which are separated by semicolons. If one such explanation takes one of the forms:

- of $noun_1$
- of $noun_1$ and $noun_2$
- of $noun_1$ or $noun_2$

then $noun_1$ and $noun_2$ respectively are objects of that verb.

Heuristic HR2 applies 134 times in WordNet, providing objects for such verbs as **generalize**, **exfoliate** or **laicize**.

The noun **company** is recognized as an object of the synset {**take over, buy out**}, and so is **corporation**. Both of them are hyponyms of {**business, concern, business concern**}, which fills in the object role of {**business, concern, business concern**}. Because of that, both **company** and **corporation** from the gloss of {**take over, buy out**} are disambiguated and point to their first corresponding synsets. Due to the inheritance property, **company** is an object of any hypernyms of {**take over, buy out**}. One such hypernym, {**buy, purchase, take**} also meets the requirements of HR3:

Heuristic Rule 3 (HR3) If a verb concept has another verb at the beginning of its gloss, then that verb describes the same action, but in a more specific context.

Therefore, **acquire** is a definition of {**buy, purchase, take**}, that has **company** as an object and involves a financial transaction. These three heuristics operate throughout all the sequences of the class comprising $<$ *acquisition of company* $>$, $<$ *addition of business* $>$, $<$ *formation of group* $>$ or $<$ *beginning of service* $>$

We conclude that for this class of prepositional relations, $noun_2$ is the object of the action described by $noun_1$.

4 A case study

Table 1 illustrates the semantic relations observed in WordNet for some of the classes of prepositional relations with preposition **of**, when both arguments are nouns. We applied a number of 28 heuristics on 45 disambiguated classes.

5 Conclusions

This paper proposes a method of extracting and validating semantic relations for prepositional attachment. The method is appealing because it uses WordNet (which is publicly available and applicable to broad English) and is scalable. A plausible explanation of prepositional attachment may be provided and the lexical disambiguation of the phrase heads is possible. The method may be improved by using additional attachment locations as provided by the transformations proposed in (Brill and Resnik, 1994).

References

Eric Brill and Philip Resnik. 1994. A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation. In *Proceedings of COLING-94*.

George Miller and Daniel Teibel. 1991. A proposal for lexical disambiguation. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 395–399, Washington, D.C.

Philip Resnik. 1995. Disambiguating Noun Grouping with Respect to WordNet Senses. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 54–68, MIT, Cambridge,Massachusets, June.

Stefan Wermter. 1991. Integration of Semantic and Syntactic Constraints for Structural Noun Phrase Disambiguation. In *Proceedings of IJCAI-91*,pages 1486–1491.