Preparing Data for Learning Noun-Modifier Semantic Relations in Base Noun Phrases

Vivi Nastase and Stan Szpakowicz

School of Information Technology and Engineering, University of Ottawa
Ottawa, Canada
{vnastase, szpak}@site.uottawa.ca

Abstract. We present experiments in automating the assignment of semantic relations to noun-modifier pairs. We work with C5.0 and memory-based learning to test the learnability of this task. *WordNet* is our semantic resource, and *SemCor* is the corpus. We work with a list of semantic relations created, and later extended and consolidated, in our long-term research project. Our results show that automating the assignment of semantic relations is possible, depending though on the learning method.

1 Introduction

We present an experiment that is part of a project to create (automatically, using learned rules) a network of semantic relations among entities in a text. We investigate the automation of the assignment of semantic relations to pairs of entities such as coordinate or subordinate clauses, a predicate and its argument, or a noun and its modifier [2], [3]. The problem is hard at the level of base noun phrases (modifier-noun pairs). Relations among clauses or inside a clause are usually marked, however ambiguously, by overt syntactic and morphological indicators, but inside a noun phrase there are very few such indicators, seldom more telling than part-of-speech information.

An ideal semantic resource for our experiment would assign semantic markers to all heads and modifiers, using some well-established and credible public-domain ontology. In the absence of such an ontology, we categorize words by their hypernyms in *WordNet*. We also plan to run a pilot experiment that uses *Roget's Thesaurus* instead of *WordNet*.

We present the sources of training data, and their processing to fit the input requirements of C5.0 and memory-based learning (MBL). We discuss the attributes that characterize our data and the acquisition of their values from the semantic resource that we use. We describe the learning experiments and the results.

2 The List of Relations

We work with a list of semantic relations derived from three lists designed by Delisle [5] and Barker [1] separately for multi-clause sentences, clauses and noun phrases. Relation assignment is learned in a semi-automatic process, seeded with the semantics of conjunctions and prepositions. Relations in noun phrases are assigned manually until enough text has been processed to allow learning to take effect.

Delisle and Barker's lists have overlaps; for example, causality or certain temporal relationships can be detected at all three levels. The unification of the lists is a major goal of our project. For the present experiment, we have concentrated on an extended list of noun-modifier relations (NMRs). The list shown in the appendix includes paraphrases that informally define the relations and, at the same time, allow us to assign relations to the training data. The modifications to Barker's NMR list are as follows:

- the **Time** relation was split into **Time At, Time From, Time To, Time Through, Frequency, Precedence and Co-occurrence**;
- the **Location** relation was split into **Location At, Location From, Location To, Location Through**;
- the **Property** relation was split into **Property**, **Manner**, **Measure**, and **Object-Property**;
- the relations **Entailment, Detraction, Prevention, Conjunction, Disjunction, Accompaniment, Exclusion, Order** were introduced.

No completely new relations were added. Each modification can be justified by describing a syntactic transformation that turns a clause (which displays a certain semantic relation) into a noun phrase while preserving the sense of the source expression. Clearly, loss of information must occur in such a transformation. We argue, however, that this loss can be ignored in our project, which does not aim to perform very deep and subtle linguistic analyses.

3 The Data

We use two sets of data in the experiments. In contrast with earlier work in our project, when technical texts were used, these data sets come from general texts.

The first, smaller, set of data (Data1) consists of base NPs extracted from Larrick [10], noun compounds from Levi [11], and a number of NPs added to balance the data set. (We want about the same number of examples for each relation). We have manually assigned senses from *WordNet 1.6* and our semantic relations to these modifier-noun pairs. Our assignment of relations was not judged by anyone else, but we have discussed each NP, and resolved all differences. The paraphrases associated with each relation, against which we tested our choices, are quite distinct.

The second set of data (SemCorData) consists of base NPs, automatically extracted from *SemCor*, the *Brown Corpus* tagged with *WordNet 1.6* senses. We have used one of *SemCor*'s three parts: 103 texts with all open class words annotated with *WordNet* senses. Extraction is based on a very simple procedure that looks at a 4-word window for part of speech sequences:

$$PosW_1, PosW_2, PosW_3, PosW_4$$

The base NP should be $W_2 - W_3$, so the patterns where it is very likely to find a base NP respect the following restrictions:

- $PosW_1 \notin \{noun, adjective\}$
- $PosW_2 \in \{noun, adjective, adverb\}$
- $PosW_3 = noun$
- $PosW_4 \neq noun$

We will not try to balance the data set obtained from *SemCor* by adding modifier-noun pairs, even if some relations appear rarely. The set will have a natural distribution of relations, as it arises from the corpus. Manual relation assignment is under way.

3.1 Building Data Sets Using WordNet

A Perl script expands the *WordNet* sense associated with each word into a list, according to the word's part-of-speech:

- NOUNS The list contains the hypernym hierarchy above the particular sense of the word. A *WordNet* synset may have several hypernyms. The list contains all paths in the hierarchy: from the synset that captures the sense of our noun up to the topmost level.
  Example: hypernym hierarchy above sense 1 of noun *person*, and the list:

```
person, individual, someone, somebody, mortal, human, soul
       => life form, organism, being, living thing
           => entity, something
       => causal agent, cause, causal agency
           => entity, something

[ [[person, individual, someone, somebody, mortal, human, soul],
   [life form, organism, being, living thing],
   [entity, something]],
  [[person, individual, someone, somebody, mortal, human, soul],
   [causal agent, cause, causal agency],
   [entity, something]]  ]
```

- ADJECTIVES, ADVERBS 1 If the word does not pertain to or derive from another word, then we take the head of the synset for adjectives, and the whole synset for adverbs.
  Example: sense 1 of adjective *light*, and the list:

```
light (vs. heavy)
       => airy                    [light, (vs.heavy)]
       => ...
```

  Example: sense 1 of adverb *daily*, and the list:

```
daily, every day, each day        [daily, every day, each day]
```

– ADJECTIVES, ADVERBS 2 If the word derives from or pertains to another word W, then we indicate a change of part of speech, and we proceed to expand the word W (according to its part of speech).
Example: sense 1 of adjective *parental*, its hypernym, and the list:

```
parental (vs. filial)
       Pertains to noun parent (Sense 1)
      =>parent
       => genitor

parent
       => genitor
           => progenitor, primogenitor
               => ...

[ [[parental,(vs.filial)],
   [parent], [genitor], [progenitor, primogenitor],
   ... , [life form, organism, being, living thing],
   [entity, something]],
  [[parental,(vs.filial)],
   [parent], [genitor], [progenitor, primogenitor],
   ... , [causal agent, cause, causal agency],
   [entity, something]] ]
```

To fit the format requirements imposed by the machine learning system C5.0, the data is post-processed by two Prolog scripts, as described in the following subsection.

The MBL process accepts the data format before postprocessing.

One could ask why we limit *WordNet*'s contribution to hypernyms. We have briefly investigated the relevance of other relations in *WordNet* and decided against using them, as they did not bring any improvement to the task at hand. A comprehensive study before definitely dismissing the additional information that *WordNet* provides is left as future work.

3.2 The Attributes

One record in the data set will contain the following attributes for the head and for the modifier: the root form, the part of speech, the source (deverbal/denominal/true adjective or adverb, deverbal/true noun), and each synset along a path in the hypernym hierarchy above a particular *WordNet* sense.

Two problems with formatting the data sets as input to C5.0 were solved by scripts written in Prolog. The first problem arises from the organization of *WordNet*. As we saw in section 3.1, a synset can have several hypernym synsets. We extract all possible paths in the hypernym hierarchy. Each modifier-noun pair is duplicated, so that only one path in the hypernym hierarchy appears in one record in the data set. (From Data1, which contained 600 modifier-noun pairs, we obtained a data set Data2 with 767 records).

The second problem is caused by *WordNet* paths having different lengths, while C5.0 expects records of equal length. We have aligned the paths in *WordNet* at the most general hypernym. We calculate the maximum length of a path for modifiers and heads, and we pad the list of each modifier and head from the left with the least general synset, to reach the maximum length.

The source of a head or modifer comes from a semi-automatically extracted list of deverbal nouns and deverbal adjectives. This attribute, one of the most significant in the identification of some relations, is included because many noun phrases arise from transformations of the verb in a clause (the arguments may undergo some morphological transformations, as well). Depending on the words that are involved, noun phrases displaying different relations will be obtained. For example:

The students protested against ...

By nominalizing the verb *protested*, we obtain the noun phrase

student protest

Verb nominalization is very productive [11]. When the modifier is the **Agent** in the original clause, it will be the **Agent** in the resulting noun phrase. When the modifier is the **Object** in the original clause, as in the example:

They analyze dreams

with the corresponding noun phrase

dream analysis

the relation in this noun phrase is also **Object**.

The problem is whether the nominalized verb does indeed preserve its relations with the arguments after transformation. Some work in this area has approached the problem from several different directions. Rappaport Hovav and Levin [8] discuss when '-er' nominals (e.g *taker, maker*) inherit the argument structure of the verb they are derived from. Macleod [12], Myers [13] take a grammatical perspective. They consider the mutual position of the arguments of the verb and the modifiers of the verb's nominal form. Hull, Gomez [9] and Gomez et al. [7] concentrate on distinguishing verbal and non-verbal senses of a nominalization, and then assigning proper thematic roles for the modifiers of the nominal, according to the underlying verb. The unstated assumption in all this work is that the nominalized verb preserves its relations with the arguments.

Other types of transformations can lead to a noun phrase that displays the **Agent** or the **Object** relation. For example:

student inventions (AGT)

(obtained by *patient* nominalization from *the inventions that the students came up with*)

blood donor (OBJ)

(obtained by *agent* nominalization from *the one who donates blood*) [11].

The verb can become either the head or the modifier in the noun phrase obtained after the transformation. For example:

They repair engines

could be paraphrased as *(They are involved in) engine repair*. The verb *repair* is nominalized, and the noun phrase has the **Object** relation. On the other hand,

They repaired the engine
could be paraphrased as *(the) repaired engine*. The verb *repair* is adjectivalized (past participle). The relation between the modifier and the head in this noun phrase is different. We have named it **Object-Property** (it identifies the object by a property that is derived from an action).

Our assumption is that noun phrases produced by these transformations preserve the meaning of the source clause, but this is not linguistically accurate. We only claim some form of common-sense meaning preservation that could be established by a non-specialist. On the other hand, some of those who study verb nominalizations actually start from the premise that transforming a clause by nominalizing the verb preserves the meaning. Though we also study the transformations that lead from one syntactic level to another, and their impact on the semantic relations, we will not further justify our working assumption in this paper.

4 The Learning Process

We use two learning algorithms, decision tree learning (C5.0) and memory-based learning.

4.1 C5.0

In Data2, the input data for our experiment, the distribution of relations was the following (Relation, Number of occurrences in our data set):

Rel	Occur	Rel	Occur	Rel	Occur	Rel	Occur	Rel	Occur
cs	19	tthr	6	ben	11	posr	43	cntr	3
eff	37	dir	8	obj	45	part	15	top	54
prp	44	loc	7	obj-prop	15	whl	10	meas	31
detr	4	lat	24	inst	44	prod	20	eq	17
freq	17	lfr	28	st	11	src	21	type	16
tat	31	ag	73	prop	52	cont	17	mat	44

We conducted one experiment with C5.0 on this data set. The target attribute was the semantic relation with 30 possible values. The tool did not handle this very well, because it built a very large decision tree. For the 767 cases, the tree had 498 leaves.

We have decided to build separate decision trees for each of the 30 relations represented in the data set. Identical copies of the data set serve as input data for learning each relation. The examples of the relation to learn are marked positive, the others negative. Each file is balanced to have a close number of positive and negative relations. We have chosen a ratio of 1:2 positive/negative examples: we want to capture as many negative examples as possible in order to properly distinguish the attributes that characterize only the positive examples, and also have enough positive examples to reinforce the choice of significant classifying attributes.

4.2 The Results Obtained with C5.0

We have conducted three-fold cross-validation experiments using C5.0 on the data set Data2 (the number of cross-validations performed is a system parameter). Partitioning our small data sets into more than 3 subsets would cause the testing data to have very few or no positive examples. We present some of the results obtained:

Relation	Set Size	Tree Size	Nr. Rules	False Positive	False Negative	Error
ag	219	18.7	15.3	5.33 (7.3%)	3 (4.1%)	11.4%
obj	135	4.3	4.3	6.33 (14.1%)	2 (4.44%)	18.5%
obj-prop	45	4	3.7	0 (0%)	0 (0%)	0%
meas	93	10.3	5.3	2 (6.45%)	5 (16.1%)	22.6%
freq	51	6.0	5.3	1.33 (7.82%)	2.67 (15.7%)	23.6%
src	63	8.0	6.0	1.67 (7.95%)	2.67 (12.7%)	20.6%
tat	93	6.3	3.7	0.33 (1.06%)	7 (22.5%)	23.6%
others						33.3%
baseline						33.3%

For most of the relations, the system with its default settings gave us the baseline error (which in our case is the error of classifying everything as negative). We have boosted the positive examples by adding a misclassification cost - we have set the cost of misclassifying a relation as negative to 2 (to compensate for the 2:1 negative/positive ratio)[1]. We noted significant improvent. Here are selected results (we omitted any result similar to one of these):

Relation	Set Size	Tree Size	Nr. Rules	False Positive	False Negative	Error
ag	219	9.3	6.7	5.67 (7.77%)	2.67 (3.66%)	11.4%
obj	135	4.3	4.3	6.33 (14.1%)	2 (4.44%)	18.5%
obj prop	45	4.0	3.0	0 (0%)	0 (0%)	0%
meas	93	13.3	6.7	4 (12.9%)	0.33 (1.1%)	14.4%
freq	51	4.0	4	1.67 (9.82%)	0.33 (1.94%)	11.8%
src	63	10.7	7.7	2.67 (12.7%)	0.67 (3.2%)	15.9%
tat	93	26.7	12.7	1 (3.2%)	1.67 (5.38%)	8.6%
cs	57	18.0	7.0	1 (5.2%)	1.67 (8.78%)	14.0%
eff	111	21.0	11.7	4.33 (11.7%)	7 (18.91%)	30.6%
mat	132	29.0	13.0	4.67 (10.6%)	3.33 (7.56%)	18.2%
loc	21	8.0	1.3	0 (0%)	1 (14.28%)	14.3%
baseline						33.3%

False Positive and *False Negative* represent the average over the 3 test runs, on the testing data, of the misclassified examples.

The error rates and numbers in the tables do not tell the whole story. For example **Location (loc)**, from the three trials of cross validation, two gave 0% error. The size of the decision tree can also be misleading, because for any choice of attribute, C5.0 builds a leaf for each of its values, and shows how each

[1] We have conducted experiments with the entire data set, keeping all the negative examples and setting the cost proportionally, but because of the great imbalance between the positive and negative classes, the results obtained were poor.

attribute value classifies the examples. In most of our cases, 2 or 3 of the values of the attributes are enough to classify the relation correctly, all the other values show negative examples.

4.3 Memory-based learning

The MBL process starts with the data described in section 3.1. Learning consists only in recording the available instances. Testing consists in assigning relations to unseen data. This is accomplished by computing distances between the test data and all the recorded instances. The relation of the example closest to our test instance will be the resulting relation [4].

In our first attempt at MBL the distance between two examples was computed using a very simple formula. The examples were represented as:
$[root_{mod}, POS_{mod}, source_{mod}, WNsense_{mod},$
$root_{head}, POS_{head}, source_{head}, WNsense_{head}]$
The formula to compute the distance between examples i and j is:

$$Dist(i,j) = \sum_k dist(a_{ik}, a_{jk})$$

$$dist(a_{ik}, a_{jk}) = \begin{cases} WNdist(a_{ik}, a_{jk}) & : & a_{ik} = root_{mod}, a_{jk} = root_{head} \\ 0 & : & a_{ik} = a_{jk} \\ 1 & : & a_{ik} \neq a_{jk} \end{cases}$$

where the *WordNet* distance $WNdist(w_1, w_2)$ is the length of the path that connects the two synsets that the words belong to, following hypernym links.

The results obtained are quite ambiguous. This learning process will assign a list of possible relations to each test example, according to the examples in the data set that are at the same distance from the test example. The MBL process does not perform well on our examples because some attributes are more significant than others. We do not know which of them are, so we cannot adapt the distance formula to give more important attributes more weight. We prefer a learning tool such as C5.0, which identifies the more relevant attributes.

5 Conclusions and Future Work

We have found that learning semantic relation assignment is possible. The relations that are marked by syntactic means (even rudimentary), **Agent, Object, Object Property**, are more easily identified. We can derive them from clause-level counterparts by applying specific transformations. The attributes used in classification are the source of the head word (for **Agent, Object**) and the source of the modifier for **Object-Property** (in particular, the modifier is the past participle of a verb).

For relations marked only by semantics (when the choice of a relation relies only on the meaning of the word), learnability depends on the ontology used. We will experiment with *Roget's Thesaurus* to test the usefulness of its ontology for our purpose.

We have observed that the deepest level of the *WordNet* hypernym hierarchy used by C5.0 was 3 (two levels down from the top). This is a surprising result and we will look for an explanation.

C5.0 does not combine or compare attributes. In order to properly learn certain relations (e.g. **Type**) this is necessary. Further work will include experimenting with other ML tools.

Longer-term plans include work on noun phrases larger than modifier-noun pairs.

6 Acknowledgements

We thank Sylvain Delisle and Ken Barker for their comments on the paper. Partial funding for this work comes from the Natural Sciences and Engineering Research Council of Canada.

References

1. K. Barker . "Semi-Automatic Recognition of Semantic Relationships in English Technical Texts", PhD Thesis, University of Ottawa, 1998.
2. K. Barker, S.Szpakowicz . "Semi-Automatic Recognition of Noun-Modifier Relationships", *Proc. COLING - ACL'98* Montreal, Canada, August, 1998, 96-102.
3. K. Barker, T. Copeck, S. Delisle, S. Szpakowicz . "Systematic Construction of a Versatile Case System", *Journal of Natural Language Engineering* 3(4), December, 1997, 279-315.
4. T. Cover, P. Hart . "Nearest Neighbor Pattern Classification",*IEEE Transactions on Information Theory*, 13, 1967, 21-27.
5. S. Delisle . "Text Processing without A-Priori Domain Knowledge: Semi-Automatic Linguistic Analysis for Incremental Knowledge Acquisition", Ph.D. thesis, TR-94-02, Department of Computer Science, University of Ottawa, 1994.
6. F. Gomez . "Linking WordNet Verb Classes to Semantic Interpretation", Technical report UCF-CS-TR-97-03, Univ. of Central Florida, 1997.
7. F. Gomez, R. Hull, C. Segami . "Acquiring Knowledge from Encyclopedic Texts", *Proc. ACL 4th Conference on Applied Natural Language Processing*, Stuttgart, Germany, 1994, 84-90.
8. M. Rappaport Hovav, B. Levin . "-er Nominals: Implications for the Theory of Argument Structure", *Syntax and Semantics* vol. 26, 1992, 127-153.
9. R. D. Hull, F. Gomez . "Semantic Interpretation of Nominalizations", *Proc. Thirteenth National Conference on Artificial Intelligence*, Portland, Oregon, August, 1996, 1062-1068.
10. N. Larrick . *Junior Science Book of Rain, Hail, Sleet and Snow*, Champaign, Garrard Publishing Company, 1961.
11. J. N. Levi . *The Syntax and Semantics of Complex Nominals*, New York, Academic Press, 1978.
12. C. Macleod, R. Grishman, A. Myers, L. Barrett, R. Reeves . "NOMLEX: A Lexicon of Nominalizations", *Proc. EURALEX'98*, Liege, Belgium, August 1998.
13. A. Myers, C. Macleod, R. Yangarber, R. Grishman, L. Barrett, R. Reeves . "Using NOMLEX to Produce Nominalization Patterns for Information Extraction", *Coling-ACL'98 Workshop Proc.: The Computational Treatment of Nominals* Montreal, Canada, August, 1998.

Appendix : Noun-Modifier Relations (NMRs)

The relations not on Barker's NMRs list are marked with a bullet (●). The relations that were not represented in our data sets are marked with a star(⋆).

Relation	Abr.	Example	Paraphrase
CAUSALITY			
cause	cs	flu virus	H makes M occur or exist, H - necess. and suff.
effect	eff	exam anxiety	M makes H occur or exist, M - necess. and suff.
purpose	prp	concert hall	H is for V-ing M, M - not necess. occurs or exists
●⋆entailment	ent		H makes M occur or exist, H - not known to exist
●detraction	detr	headache pill	H opposes M, H - not suff. to prevent M
●⋆prevention	prev		H opposes M, H - suff. to prevent M
TEMPORALITY			
●⋆co-occurrence	cooc		H and M occur or exist at the same time
●frequency	freq	daily exercise	H occurs every time M occurs
●⋆precedence	prec		H (begins to) occurs or exists before M
time at	tat	morning exerise	H occurs when M occurs
●⋆time from	tfr		H began to occur when M became true
●time through	tthr	six-hour meeting	H existed while M existed, M - interval of time
●⋆time to	tto		H existed until M started to exist
SPATIAL			
●direction	dir	outgoing mail	H is directed towards M, M is not the final point
location	loc	home town	H is the location of M
location at	lat	desert storm	H is located at M
●location from	lfr	foreign capital	H originates at M
●⋆location to	lto		the destination of H is M
●⋆location through	lthr		H occurred through M (M is a space)
CONJUNCTIVE			
●⋆conjunction	conj		both H and M exist
●⋆disjunction	disj		either one or both H and M exist
PARTICIPANT			
●⋆accompaniment	acc		H is accompanied by M (co-agent)
agent	ag	student protest	M perfomrs H, M - animate or natural phen.
beneficiary	ben	student discount	M benefits from H
●exclusion	excl		M is excluded from H, or H replaces M
instrument	inst	laser printer	H uses M
object	obj	metal separator	M is acted upon by H
●object property	obj prop	sunken ship	H underwent M
part	part	printer tray	H is part of M
posessor	posr	national debt	M has H
property	prop	blue book	H is M
product	prod	plum tree	H produces M
source	src	olive oil	M is the source of H
stative	st	sleeping dog	H is in a state of M
whole	whl	daisy chain	M is part of H
QUALITY			
container	cntr	film music	M contains H
content	cont	apple cake	M is contained in H
equative	eq	player coach	H is also M
manner	man	stylish writing	H occurs in the way indicated by M
material	mat	brick house	H is made of M
measure	meas	expensive book	M is a measure of H
order	ord		H is before M in physical space
topic	top	weather report	H is concerned with M
●type	type	oak tree	M is a type of H