Large Scale Experiments for Semantic Labeling of Noun Phrases in Raw Text

Louise Guthrie[*], **Roberto Basili**[†], **Fabio Zanzotto**[†], **Kalina Bontcheva**[*],
Hamish Cunningham[*], **David Guthrie**[*], **Jia Cui**[**], **Marco Cammisa**[†],
Jerry Cheng-Chieh Liu[†††], **Cassia Farria Martin**[****], **Kristiyan Haralambiev**[††††],
Martin Holub[***], **Klaus Macherey**[††], **Fredrick Jelinek**[**]

[*]University of Sheffield [**]The Johns Hopkins University [***]Charles University [****]Harvard University
[†]University of Rome Tor Vergata [††]University of Aachen [†††]Colulmbia University [††††]University of Bulgaria

Abstract

This paper gives a brief overview of the results of our work during the Summer 2003 Workshop of the Center for Language and Speech Processing at the Johns Hopkins University in Baltimore Maryland. The goal of the project was to determine the feasibility of extending named entity recognition to common nouns and determine whether or not it is possible to assign automatically a predetermined set of semantic tags and approach human performance in the task.

1. Introduction

Although it is generally assumed that improvements in language processing will be made through the integration of linguistic information and statistical techniques, the reality is that language is very diverse and looking for specific patterns of words that repeat enough to be statistically significant tends not to be a very fruitful task: sequences longer than three words are not generally repeated often enough to be statistically significant. At the same time, the identification of named entities: Names, dates, places, organizations etc., has proved to be a very useful preliminary task in many natural language processing systems(Appelt, 1999; Grishman and Sundheim, 1996; Cunningham, 1999), in some sense, because it attacks the data sparseness problem by collapsing (semantically) related phrases which are expressed by different word sequences. This project extends that notion to common nouns not marked as a named entity and we hope contributes to the goal of allowing repeatable semantic patterns to emerge from text.

2. Goal of the work

The project investigated the feasibility of automatically and accurately assigning coarse-grained semantic labels to noun phrases in a 26 million word subset of the British National Corpus (BNC). The labels were chosen because they are a small set that gave us relatively good coverage in a general corpus. The focus of the project, however, is not to defend the particular set of labels but rather to suggest a way, given a set of labels, to extend named entity recognition to common noun phrases (other than time and currency references).

3. The Tag Set

The particular set of 21 semantic labels used in the experiments was inspired by the semantic categories assigned to noun senses in the electronic version of Longman's Dictionary of Contemporary English (LDOCE)(Procter, 1978).

The categories in LDOCE were used to specify the semantic category of a noun, as well as to indicate adjective preferences (the kind of noun the adjective likes to modify), and the verb preferences (the kind of subject, object and indirect object the verb expects). Although not all of the senses for nouns, verbs, and adjectives contain these markings in the dictionary, the vast majority of senses do, and as this project was interested in the assignment of a small, predetermined set of tags to a large general corpus, this seemed an appropriate choice for a set of pre-defined tags.

The LDOCE tag set was modified somewhat for our experiments. Several of the semantic categories in LDOCE were used to describe disjunctive or conjunctive categories (for example abstract or solid). These were not included in the 21 semantic tags used in the experiments. It should be noted that the particular set of tags chosen did not include any subdivisions of abstract nouns. Most of the semantic categories are sub-divisions of concrete nouns, so the task was to label abstract noun phrases as abstract, and otherwise to assign one of the other 20 semantic categories. The semantic classes used in the experiments include those in Figure 1 below as well as:

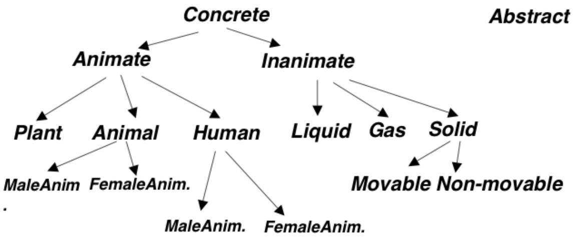


Figure 1: Some of the Semantic categories

- Collective - for animals or collective for humans
- Physical Qualities - refers to physical qualities not associated with actual matter (diseases, measurements of physical dimensions, heat, light, etc...)
- Organic Material - refers to things which form part of a living organism such as to bone, tissue, bark, etc.
- Male - either male animal or male human
- Female - either female animal or human

4. The Corpus

The corpus consists of approximately 26 million words of written data from the British National Corpus (BNC). This subset of the BNC contains documents from the domains of science, social science, world affairs, and business:

4.1. The Corpus Annotation

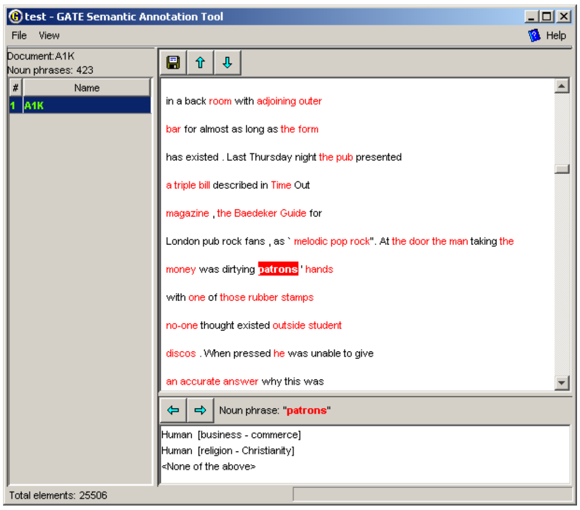
The corpus annotations for the entire 26 million word corpus are extensive, and are documented in the final report which can be found on the web page for the project. (http://www.clsp.jhu.edu/ws03/groups/sparse/). The corpus annotations, the annotation tool (section 4.2.), and many of the experiments were developed as part of GATE (Cunningham et al., 1997; Bontcheva et al., 2002), a General Architecture for Text Engineering. The system is now a widely-used and relatively comprehensive infrastructure for language processing software development. Details are available at (http://gate.ac.uk). The annotations include:

- Named entities: *Date*, *Location*, *Person*, *Organization*, *Time* and *Currencies* are recognized using the MUSE(Maynard et al., 2003) NER.

- The remaining basic noun phrases are recognized and head nouns are identified (they are assumed to be the last noun in the noun phrase).

4.2. The Annotation Tool

The corpus annotation tool takes pre-processed documents, collected in a corpus, and provides the human annotators with an intuitive, fast interface, which enables them to annotate nouns by choosing from a list of valid semantic tags. The documents are in GATE stand-off XML format and already contain the nouns that need to be marked and all the possible semantic tags that they have in the LDOCE lexicon. Subsequent to the annotation process, the corpus was converted into inline XML, so it can processed easily with script languages, such as perl. Figure 2 shows the tool in action.



Document:A1K
Noun phrases: 423

in a back room with adjoining outer
bar for almost as long as the form
has existed . Last Thursday night the pub presented
a triple bill described in Time Out
magazine , the Baedeker Guide for
London pub rock fans , as ` melodic pop rock". At the door the man taking the
money was dirtying patrons ' hands
with one of those rubber stamps
no-one thought existed outside student
discos . When pressed he was unable to give
an accurate answer why this was

Noun phrase: "patrons"

Human [business - commerce]
Human [religion - Christianity]
<None of the above>

Total elements: 25506

Figure 2: A screen shot of the annotation tool

4.3. The Human annotated data

Human annotators (native speaking PhD students in computer science) were asked to choose the correct semantic category for the head nouns of basic noun phrases in a one million word set of sample documents taken from our corpus. The annotators were asked to choose from among the possible semantic categories for those head nouns provided that they were also found in LDOCE and choose ’none of the above’ if the noun was used in a way that did not correspond to any of the semantic categories in their list. The annotations produced 214,446 instances where a specific semantic category was chosen and these instances comprised the human annotated data for the experiments.

4.4. Double-Annotated Data

A portion of the corpus was double-annotated, in order to allow analysis of the extent to which the human annotators agreed (or *inter-annotator agreement*).

The total number of instances (headwords) that were marked by both annotators (where neither annotator chose ’none of the above’) was 8446. Counting only instances where the annotators matched exactly in the chosen category, the inter-annotator agreement was 94%.

5. The Experimental set-up

5.1. The Blind Data

Ten percent of the human annotated data (which contained 13, 097 instances of annotated noun phrases) was reserved for blind testing at the very end of the workshop. The remainder of the human annotated data was used to create several development corpora. In all cases, the development corpus was divided into two sections: one for training and one for testing (held out). We measured our results and honed our systems on the results obtained from these development sets and then ran the best systems over the blind data at the end of the workshop.

5.2. The Development Sets

From the start of the experiments, we wanted to train models to deal with words not seen in training, and examine the use of context on unseen words.

We created a development set (Hard) in which the held out data contained semantically ambiguous instances (noun phrases where the head noun had senses with 2 or more semantic categories in LDOCE) that never appeared in the training data. This set was used to help see how well the techniques would perform on unseen data which was ambiguous (often highly so) when it had never been seen in the training data. It consisted of 125,987 instances of unambiguous words for training and 73,371 ambiguous words in the held-out set.

The main development set (DEV) was constructed in two steps: First, a random selection of instances (85,000) were selected for the held out portion and the rest of the data was put into the training portion. Next, all instances of a few ambiguous (often highly ambiguous) words were removed from the training portion and put into the held-out data. The resulting development set consisted of 98,384 instances used for training and 100,974 instances held out to test our techniques.

6. Experiments

Although we conducted experiments that did not make use of the training data (unsupervised), no discussion of this work will be reported here.

6.1. The Baselines

The Baseline was measured by assigning the tag most frequently observed for that word if it appeared in training, and to assign the most frequent category (abstract) if the word had not been seen in training.

The baseline for the Hard development set was 45%, and the baseline for our final development set was 80%.

6.2. Bag of Words Experiments

The bag of words module was developed to comply with the multiple-knowledge sources WSD architecture (Stevenson and Wilks, 2001). The idea is to enable the use of multiple taggers and combine their results through a weighted function and Stevenson shows how such weights can be learned from a corpus.

In this work, all taggers were implemented as GATE components and the Bag-of-Words (BoW) tagger is an Information Retrieval-inspired tagger with parameters:

- Window size: 50 default value. The number of words to the left and right of the current word, which will be included in the content vector.
- What part of speech to put in the content vectors (default: nouns and verbs).
- Whether to restrict the possible categories for each word, according to the LDOCE dictionary.

The algorithm uses a list of stop words to eliminate frequent words, which bear little content and might skew the similarity measure.

The content vector approach was used in word sense disambiguation by (Leacock et al., 1993). A similar approach is used here in the training stage, to construct a context vector for each semantic category of a word. For example:

```
Crane/Animal={species, captivity, disease}
Crane/Mov.Solid={worker, disaster, machinery}
```

Seen words are classified by calculating the inner product between their context vector and the vectors for each possible category for that word.

Inner product is calculated in two ways: using binary vectors number of matching terms, and using weighted vectors with Leacock's measure (which favours concepts that occur frequently in exactly one category.)

The combined architecture is best at 93.2% (window size 50, using only nouns, binary vectors), because it uses the baseline frequency tagger to assign semantic categories to words that have not been encountered in the training data and the BoW cannot resolve, without corresponding context vectors.

6.3. Maximum Entropy Methods

Many experiments were conducted using machine learning techniques within the YASMET(Och, 2002), JME(Cui, 2003), and WEKA(Witten and Frank, 1999) toolkits. The challenge was to define feature functions that describe the corpus information we gathered, information from outside resources (a Dictionary or WordNet or another corpus), and the syntactic information gained from parsing or pattern matching. We do not give formal definitions of these feature functions in this paper (Formal definitions can be found in (Guthrie and et al., 2003)), but rather try to summarize the kind of information that was encoded as features in the machine learning toolkits. The table below show the results for the DEV set and Blind Data. Results were obtained using YASMET toolkit with a variety of feature functions (although very similar results were obtained using the JME toolkit), unless otherwise indicated.

Table 1: Development Corpus Experiments

Dev-Corpus Result Summary	
Experiment	TA [%]
Baseline	80.2
Bag of Words	81.1
f_{LEX}	80.5
$f_{\text{L-PREF}} + f_{\text{pruLongADJ}}$	84.5
$f_{\text{L-PREF}} + f_{\text{pruCorpusSDJ}}$	84.9
$f_{\text{LEX}} + f_{\text{LONG-PREF}}$	84.6
$f_{\text{LEX}} + f_{\text{LONG-PREF}}$ (lemma)	85.6
Parsing Context (VS, VO, VPP, NPP)	83.9
Unsupervised approach using WordNet	70.4

Table 2: Blind Corpus Experiments

BLIND Result Summary	
Experiment	TA [%]
Baseline	90
Bag of Words	93.2
f_{LEX}	93
$f_{\text{L-PREF}} + f_{\text{pruLongADJ}}$	88.3
$f_{\text{L-PREF}} + f_{\text{pruCorpusADJ}}$	91.9
$f_{\text{LEX}} + f_{\text{LONG-PREF}}$	92.4
$f_{\text{LEX}} + f_{\text{LONG-PREF}}$ (lemma)	92.2
Unsupervised approach using WordNet	75.2

6.4. Intuition behind Features

- f_{LEX}- This is similar to what was used in the Bag of words experiments, (where the system always choses among the possible semantic categories for the particular word, as opposed to choosing from the complete set of 21 semantic categories). In the Maximum entropy framework, nothing is precluded, but 'desirable' semantic categories (one of the semantic categories associated with the senses of that word in LDOCE are encoded and used as a feature.

- $f_{\text{LONG-PREF}}$- This features is a weighted version of the feature above and makes some use of the subject area codes available in LDOCE, together with the possible semantic categories. A reduced sense list is formed from the dictionary which only distinguishes senses if they have a distinct (subject code, semantic category) pair. For a word w, one semantic category is encoded 'more desirable' than another if it appears in more of the pairs corresponding to w.

- $f_{\text{pruLongADJ}}$- Adjective preferences (semantic categories of nouns that a given adjective tends to modify) are given for many adjectives in LDOCE. Before this feature function was encoded, all adjectives which 'strongly prefer' a given category were clustered. In this feature function, information is encoded to describe whether or not any of the modifiers of the the head noun in the noun phrase to be tagged, belong to one of the predefined adjective classes.

- $f_{\text{pruCorpusADJ}}$- This feature is similar to the one above, but the adjective clusters were not defined by dictionary preferences. Statistics of the semantic categories that adjectives modify were gathered from a large external corpus and used to select adjectives which are predictive of semantic categories. We used the remaining 75 million words of the BNC in the following way: All noun phrases were identified and a sub corpus was created of all noun phrases whose head noun was unambiguous (meaning only one possible semantic category in LDOCE). This gave a set of instances of adjectives modifying classes of nouns. Before this feature function was encoded, all adjectives which 'strongly prefer' a given category were clustered. 'Strongly prefer' was measured by computing the entropy of the adjective with respect to the semantic classes. Low entropy indicates the adjective is a good distinguisher of class, whereas high entropy means it is not.

- Parsing Context (VS, VO, VPP, NPP)- These experiments used the JME toolkit, and encoded as features verbs that preferred a particular class of subject, a particular class of object and a particular head noun in a modifying prepositional phrase, they also considered prepositions that preferred a certain class of noun. Description of these is given in the project report.

- Unsupervised approach using WordNet- Although this work will not be reported in this paper, we indicate our best results when no training data is used.

7. Conclusion

Our experiments showed that contextual information is important to the task. Simple bag of words techniques performed similarly to the maximum entropy method on the blind data that was randomly selected, however it performed worse on our DEV set, which means that words never seen in the training material might be better handled with the maximum entropy method. The inclusion of the syntactic information seemed to have little effect on the experiment, although some gains were achieved by clustering syntactic information. In the case of parsing data, results are still too sparse to create effective clusters, and further work should be done here on a bigger scale. Overall, we conclude that the development of automatic methods to accurately assign coarse semantic labels to common noun phrases is indeed feasible with an accuracy approaching that of named entity recognition.[1]

8. References

Appelt, D., 1999. An Introduction to Information Extraction. *Artificial Intelligence Communications*, 12(3):161–172.

Bontcheva, K., H. Cunningham, V. Tablan, D. Maynard, and H. Saggion, 2002. Developing Reusable and Robust Language Processing Components for Information Systems using GATE. In *3rd International Workshop on Natural Language and Information Systems (NLIS'2002)*. Aix-en-Provence, France: IEEE Computer Society Press.

Cui, Jia, 2003. Jia's Maximum Entropy Toolkit. JME http://www.clsp.jhu.edu.

Cunningham, H., 1999. Information Extraction: a User Guide (revised version). Research Memorandum CS–99–07, Department of Computer Science, University of Sheffield.

Cunningham, H., K. Humphreys, R. Gaizauskas, and Y. Wilks, 1997. Software Infrastructure for Natural Language Processing. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*.

Grishman, R. and B. Sundheim, 1996. Message understanding conference - 6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen.

Guthrie and Basili et al., 2003. Semantic Analysis for Data Sparsity Compensation. Final Report http://www.clsp.jhu.edu/ws2003/groups/sparse/.

Leacock, C., G. Towell, and E. Voorhees, 1993. Corpus-Based Statistical Sense Resolution. In *Proceedings of ARPA Human Language Technology Workshop*.

Maynard, D., V. Tablan, K. Bontcheva, H. Cunningham, and Y.Wilks, 2003. Multi-source entity recognition – an information extraction system for diverse text types. Research Memorandum CS–03–02, Department of Computer Science, University of Sheffield.

Och, Franz Josef, 2002. Yasmet. http://www-i6.Informatik.RWTH-Aachen.de/web/Software/index.html.

Procter, P., 1978. *Longman Dictionary of Contemporary English*. Essex: Longman Group.

Stevenson, M. and Y. Wilks, 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*.

Witten, I. H. and E. Frank, 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

[1] This material is based upon work supported by the National Science Foundation under Grant No. 0121285.