# SEXTANT: EXPLORING UNEXPLORED CONTEXTS FOR SEMANTIC EXTRACTION FROM SYNTACTIC ANALYSIS

**Gregory Grefenstette**
*Computer Science Department, University of Pittsburgh, Pittsburgh, PA 15260*
grefen@cs.pitt.edu

## Abstract

For a very long time, it has been considered that the only way of automatically extracting similar groups of words from a text collection for which no semantic information exists is to use document co-occurrence data. But, with robust syntactic parsers that are becoming more frequently available, syntactically recognizable phenomena about word usage can be confidently noted in large collections of texts. We present here a new system called SEXTANT which uses these parsers and the finer-grained contexts they produce to judge word similarity.

## BACKGROUND

Many machine-based approaches to term similarity, such as found in TRUMP (Jacobs and Zernick 1988) and FERRET (Mauldin 1991), can be characterized as knowledge-rich in that they presuppose that known lexical items possess Conceptual Dependence(CD)-like descriptions. Such an approach necessitates a great amount of manual encoding of semantic information and suffers from the drawbacks of **cost** (in terms of initial coding, coherence checking, maintenance after modifications, and costs derivable from a host of other software engineering concern); of **domain dependence** (a semantic structure developed for one domain would not be applicable to another. For example, *sugar* would have very different semantic relations in a medical domain than in a commodities exchange domain); and of **rigidity** (even within well-established domain, new subdomains spring up, e.g. AIDS. Can hand-coded systems keep up with new discoveries and new relations with an acceptable latency?)

In the Information Retrieval community, researchers have consistently considered that "the linguistic apparatus required for effective domain-independent analysis is not yet at hand," and have concentrated on counting document co-occurrence statistics (Peat and Willet 1991), based on the idea that words appearing in the same document must share some semantic similarity. But document co-occurrence suffers from two problems: **granularity** (every word in the document is considered potentially related to every other word, no matter what the distance between them) and **co-occurrence** (for two words to be seen as similar they must physically appear in the same document. As an illustration, consider the words *tumor* and *tumour*. These words certainly share the same contexts, but would never appear in the same document.) In general different words used to describe similar concepts might not be used in the same document, and are missed by these methods.

Recently, a middle ground between these two approaches has begun to be broken. Researchers such as (Evans *et al.* 1991) and (Church and Hanks 1990) have applied robust grammars and statistical techniques over large corpora to extract interesting noun phrases and subject-verb, verb-object pairs. (Hearst 1992) has shown that certain lexical-syntactic templates can reliably extract hyponym relations froM text. (Ruge 1991) shows that modifier-head relations in noun phrases extracted from a large corpus provide a useful context for extracting similar words. The common thread of all these techniques is that they require no hand-coded domain knowledge, but they examine more cleanly defined contexts than simple document co-occurrence methods.

Similarly, our SEXTANT [1] uses fine-grained syntactically derived contexts, but derives its measures of similarity from consider-

[1] Semantic EXtraction from Text via Analyzed Networks of Terms

ing not the co-occurrence of two words in the same context, but rather the overlapping of all the contexts associated with words over an entire corpus. Calculation of the amount of shared weighted contexts produces a similarity measure between two words.

## SEXTANT

SEXTANT can be run on any English text, without any pre-coding of domain knowledge or manual editing of the text. The input text passes through the following steps: (I) Morphological analysis. Each word is morphologically analyzed and looked up in a 100,000 word dictionary to find its possible parts of speech. (II) Grammatical Disambiguation. A stochastic parser assigns one grammatical category to each word in the text. These first two steps use CLARIT programs (Evans *et al.* 1991). (III) Noun and Verb Phrase Splitting. Each sentence is divided into verb and noun phrases by a simple regular grammar. (IV) Syntagmatic Relation Extraction. A four-pass algorithm attaches modifiers to nouns, noun phrases to noun phrases and verbs to noun phrases. (Grefenstette 1992a) (V) Context Isolation. The modifying words attached to each word in the text are isolated for all nouns. Thus the context of each noun is given by all the words with which it is associated throughout the corpus. (VI) Similarity matching. Contexts are compared by using similarity measures developed in the Social Sciences, such as a weighted Jaccard measure.

As an example, consider the following sentence extracted from a medical corpus.

```
Cyclophosphamide markedly prolonged induction
time and suppressed peak titer irrespective of
the time of antigen administration.
```

Each word is looked up in a online dictionary. After grammatical ambiguities are removed by the stochastic parser, the phrase is divided into noun phrases(NP) and verb phrases(VP), giving,

```
NP      cyclophosphamide (sn)
--      markedly (adv)
VP      prolong (vt-past)
NP      induction (sn) time (sn)
--      and (cnj)
VP      suppress (vt-past)
NP      peak (sn) titer (sn) irrespective-of (prep)
        the (d) time (sn) of (prep) antigen (sn)
        administration (sn)
```

Once each sentence in the text is divided into phrases, intra- and inter-phrase structural relations are extracted. First noun phrases are scanned from left to right(NPLR), hooking up articles, adjectives and modifier nouns to their head nouns. Then, noun phrases are scanned right to left(NPRL), connecting nouns over prepositions. Then, starting from verb phrases, phrases are scanned before the verb phrase for an unconnected head which becomes the subject(VPRL), and likewise to the right of the verb for objects(VPLR), producing for the example:

```
VPRL    cyclophosphamide , prolong < SUBJ
NPRL    time , induction < NN
VPLR    prolong , time < DOBJ
VPRL    cyclophosphamide , suppress < SUBJ
NPRL    titer , peak < NN
VPLR    suppress , titer < DOBJ
NPLR    titer , time < NNPREP
NPRL    administration , antigen < NN
```

Next SEXTANT extracts a user specified set of relations that are considered as each word's context for similarity calculations. For example, one set of relations extracted by SEXTANT for the above sentence can be

```
cyclophosphamide prolong-SUBJ
time induction
time prolong-DOBJ
cyclophosphamide suppress-SUBJ
titer peak
titer suppress-DOBJ
titer time
administration antigen
time administration
```

In this example, the word *time* is found modified by the words *induction*, *prolong-DOBJ* and *administration*, while *administration* is only considered by this set of relations to be modified by *antigen*. Over the whole corpus of 160,000 words, one can consider what modifies *administration*. Isolating these modifiers gives a list such as

```
...
administration androgen
administration antigen
administration aortic
administration aramine
administration associate-DOBJ
administration associate-SUBJ
administration azathioprine
administration carbon-dioxide
administration case
administration cause-SUBJ
...
```

At this point SEXTANT compares all the other words in the corpus, using a user-specified similarity measure such the Jaccard measure, to find which words are most similar to which others. For example, the words found as most similar to *administration* in this medical corpus were the following words in order of most to least similar:

```
administration   injection, treatment, therapy,
                 infusion, dose, response, ...
```

As can be seen, the sense of *administration* as in the "administration of drugs and medicines" is clearly extracted here, since *administration* in this corpus is most similarly used as other words such as *injection* and *therapy* having to do with dispensing drugs and medicines. One of the interesting aspects of this approach, contrary to the coarse-grained document co-occurrence approach, is that *administration* and *injection* need never appear in the same document for them to be recognized as semantically similar. In the case of this corpus, *administration* and *injection* were considered similar because they shared the following modifiers:

```
acid follow-DOBJ growth prior produce-IOBJ
dose extract increase-SUBJ intravenous
treat-IOBJ associate-SUBJ associate-DOBJ
rapid cause-SUBJ antigen adrenalectomy
aortic hormone subside-IOBJ alter-IOBJ
folic-acid amd folate
```

It is hard to select any one word which would indicate that these two words were similar, but the fact that they do share so many words, and more so than other words, indicates that these words share close semantic characteristics in this corpus.

When the same procedure is run over a corpus of library science abstracts, *administration* is recognized as closest to

```
administration   graduate, office, campus,
                 education, director, ...
```

Similarly *circulation* was found to be closest to *flow* in the medical corpus and to *date* in the library corpus. *Cause* was found to be closest to *etiology* in the medical corpus and to *determinant* in the library corpus. Frequently occurring words, possessing enough context, are generally ranked by SEXTANT with words intuitively related within the defining corpus.

## DISCUSSION

While finding similar words in a corpus without any domain knowledge is interesting in itself, such a tool is practically useful in a number of areas. A lexicographer building a domain-specific dictionary would find such a tool invaluable, given a large corpus of representative text for that domain. Similarly, a Knowledge Engineer creating a natural language interface to an expert system could use this system to cull similar terminology in a field. We have shown elsewhere (Grefenstette 1992b), in an Information Retrieval setting, that expanding queries using the closest terms to query terms derived by SEXTANT can improve recall and precision. We find that one of the most interesting results from a linguistic point of view, is the possibility automatically creating corpus defined thesauri, as can be seen above in the differences between relations extracted from medical and from information science corpora. In conclusion, we feel that this fine grained approach to context extraction from large corpora, and similarity calculation employing those contexts, even using imperfect syntactic analysis tools, shows much promise for the future.

## References


(Church and Hanks 1990) K.W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), Mar 90.

(Evans *et al.* 1991) D.A. Evans, S.K. Henderson, R.G. Lefferts, and I.A. Monarch. A summary of the CLARIT project. TR CMU-LCL-91-2, Carnegie-Mellon, Nov 91.

(Grefenstette 1992a) G. Grefenstette. Sextant: Extracting semantics from raw text, implementation details. TR CS92-05, University of Pittsburgh, Feb 92.

(Grefenstette 1992b) G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. *SIGIR'92*, Copenhagen, June 21-24 1992. ACM.

(Hearst 1992) M.A. Hearst. Automatic acquisition of hyponyms from large text corpora. *COLING'92*, Nantes, France, July 92.

(Jacobs and Zernick 1988) P. S. Jacobs and U. Zernick. Acquiring lexical knowledge from text: A case study. In *Proceedings Seventh National Conference on Artificial Intelligence*, 739–744, Morgan Kaufmann.

(Mauldin 1991) M. L. Mauldin. *Conceptual Information Retrieval: A case study in adaptive parsing.* Kluwer, Norwell, 91.

(Peat and Willet 1991) H.J. Peat and P. Willet. The limitations of term co-occurrence data for query expansion in document retrieval systems. *JASIS*, 42(5), 1991.

(Ruge 1991) G. Ruge. Experiments on linguistically based term associations. In *RIAO'91*, 528–545, Barcelona, Apr 91. CID, Paris.