Use of Syntactic Context to Produce Term Association Lists for Text Retrieval

Gregory Grefenstette

Computer Science Department

University of Pittsburgh

Pittsburgh, PA 15260

grefen@cs.pitt.edu

Abstract

One aspect of world knowledge essential to information retrieval is knowing when two words are related. Knowing word relatedness allows a system given a user's query terms to retrieve relevant documents not containing those exact terms. Two words can be said to be related if they appear in the same contexts Document co-occurrence gives a measure of word relatedness that has proved to be too rough to be useful. The relatively recent apparition of on-line dictionaries and robust and rapid parsers permits the extraction of finer word contexts from large corpora. In this paper, we will describe such an extraction technique that uses only coarse syntactic analysis and no domain knowledge. This technique produces lists of words related to any word appearing in a corpus. When the closest related terms were used in query expansion of a standard information retrieval testbed, the results were much better than that given by document co-occurence techniques, and slightly better than using unexpanded queries, supporting the contention that semantically similar words were indeed extracted by this technique.

15th Ann Int'l SIGIR '92/Denmark-6/92

1 Introduction

With the current availability of machine readable dictionaries, large corpora of natural language text, and robust syntactic processors, there is a renewed interest in extracting knowledge automatically from large quantities of text. One aspect of world knowledge that is of interest for information retrieval systems is knowing when two words are related. Knowing word relatedness allows a system, given a user's query terms, to retrieve relevant documents not containing those exact terms.

As an operational definition, two words can be said to be related if they appear in the same context. Information Retrieval research from (Salton 1971) to (Peat and Willet 1991) has only considered one type of completely automatically extractable context: document co-occurrence. The document co-occurrence hypothesis is that two words appearing in the same document share some semantic relatedness. While this is certainly true, document co-occurrence is only one rough measure of a word's context. A number of papers

have called into doubt the usefulness of document co-occurrence derived similarity (Minker *et al.* 1972)(Sparck Jones 1991)(Peat and Willet 1991). 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. For example, in this paragraph *distance* and *operational* will be brought together as a data pair for the similarity measure, despite their distance.
- co-occurrence, for two words to be seen as similar they must physically appear in the same document. As a trivial counter example, consider the words *tumor* and *tumour*. These words certainly share the same contexts, but would never appear in the same document, at least not with a frequency to be recognized by any document co-occurrence method. In general different words used to describe similar concepts may not be used in the same document, and are missed by these methods.

Other Information Retrieval researchers have calculated similarity using co-occurrence in groups of documents corresponding to manual indexing categories (Lewis and Croft 1990). This approach is useful for large collections of indexed newswire, for example, where each story is headed by its news category. Another interesting technique for calculating similarity of infrequently occurring terms is described in (Crouch 1990), where documents are first clustered using a tightly clustering complete-link method and infrequent words found in the same cluster are considered similar. For the case where queries and relevancy judgements are available, (Yu and Raghavan 1977) proposed creating semantic relations between query words and words appearing in relevant documents not retrieved for that query.

In this paper, we present a technique for extracting similarity lists for words in a corpus for which no manual indexing, or relevance measures might exist. Using a simple syntactic approach for defining context, we will describe how to use these finer-grained contexts for defining relatedness. A fully implemented technique will be described which can tractably extract such contexts from large corpora. The extraction uses only coarse syntactic analysis and no domain knowledge. Syntactic patterns are used to specify the contexts of a word over the corpus. This technique, in addition to providing finer granularity of context than document co-occurence, also allows for words not appearing in the same document to be recognized as similar.

First we will describe the types of context that our system can extract. These contexts are compared using standard similarity measures, described in the third section, in the evaluation section, we show how the closest words were used in a query expansion experiment, giving the results compared with expansion via document co-occurrence data.

2 Contexts

Our basic premise is that words found in the same context tend to share semantic similarity. If we find two words which are modified by *domesticated, hairy*, and which govern the verbs *eat, drink*, and *jump*, then we would probably say that the objects that the words stand for share some similarity, even without knowing what they are. This, of course, is the position that the computer is in: not knowing what things are, but being able to recognize like contexts.

Use of syntactic analysis opens up a much wider range of contexts than simple document co-occurrence, or co-occurrence within a window of words as in (Phillips 1985). Syntactic analysis allows us to know what words modify other words, and to develop contexts from this information. We have concentrated at first on simple nouns, and the rest of the paper will Only consider similarity between words from this grammatical category.

The contexts that we recognize in our system

are

- ADJ, NN: when a word is modified by an adjective or another noun such as in

```
...most valuable player...
=> player , valuable < ADJ
...I 've included categories like
rookie pitcher...
=> pitcher , rookie < NN
```

- NNPREP: when a word is modified by a noun via a preposition such as

```
...til the end of the year...
=> end , year < NNPREP
...factors in defensive as well as
offensive performance...
=> factor , performance < NNPREP
```

- SUBJ, DOBJ, IOBJ: when a word appears as the subject, direct or indirect object of a verb. We take a simplified view of indirect objects, retaining the first prepositional phrase after a verb as its indirect object. For example:

```
...if the giants had won the nl west...
=> giant , win < SUBJ
...someone could suggest a better
formula...
=> suggest , formula < DOBJ
...reaching on an error...
=> reach , error < IOBJ
```

Since we are dealing with large corpora, and employing statistical similarity methods in which the frequency of contexts counts, the extraction of these contexts need not be perfect. A measure of error in any of the steps described below can be tolerated, if the false information it generates is limited, although we have not yet examined how much error can be introduced before performance degrades. Our technique for extracting and using these contexts follows the steps below. Each step is independent of the next, and is roughly linear-time, except for the calculation of similarities which is quadratic in the total number of unique word-context pairs extracted.

2.1 Morphological Analysis

Given a corpus, we first morphologically analyze each word in the corpus. The analysis provides the grammatical categories that every word may play. This can be performed by dictionary look-up, and/or by using morphological analysis algorithms. We employ the CLARIT (Evans *et al.* 1991a) morphological package. See Figures 1 and 2.

```
.I 1
.W
correlation between maternal and fetal
plasma levels of glucose and free
fatty acids .  correlation
coefficients have been determined
between the levels of glucose and ffa
in maternal and fetal plasma collected
at delivery .  significant
correlations were obtained between the
maternal and fetal glucose levels and
the maternal and fetal ffa levels .
from the size of the correlation
coefficients and the slopes...
```

Figure 1: Original text from corpus.

```
"correlation" sn correlation
"between" prep between
"maternal" adj maternal
"and" cnj and
"fetal" adj fetal
"plasma" sn plasma
"levels" pn level vt-pressg3 level
"of" prep of
"glucose" sn glucose
...
```

Figure 2: After morphological analysis. The original source word is followed by pairs of grammatical values and normalized words. 'levels' is ambiguous.

2.2 Syntactic Disambiguation

Next each word needs to be grammatically disambiguated. This consists of assigning a single grammatical category to each word. A number of robust grammar based or stochastic methods have been proposed (DeRose 1988) (Hindle 1989). We use a disambiguator implementing a time linear stochastic grammar based on Brown corpus frequencies, see (Evans *et al.* 1991b). See Figure 3.

2.3 Noun and Verb Phrases

Then, using a method detailed in (Grefenstette 1992) we take the disambiguated text and di-

```
...
"between" prep between
"maternal" adj maternal
"and" cnj and
"fetal" adj fetal
"plasma" sn plasma
"levels" pn level
"of" prep of
...
```

Figure 3: Each word is disambiguated by using a grammar or a simple precedence parser.

```
NP    correlation between maternal and
      fetal plasma level of glucose
      and free fatty acid
NP    correlation coefficient
VP    have be determine
NP    between the level of glucose and
      ffa in maternal and fetal plasma
VP    collect
NP    at delivery
NP    significant correlation
VP    be obtain
NP    between the maternal and fetal
      glucose level and the maternal
      and fetal ffa level
```

Figure 4: Text divided into noun and verb phrases.

vide it into verb and noun phrases. The method employs lists of grammatical values which can start and end a verb and noun phrase, and precedence matrices describing legal continuations of verb and noun phrases. Our definition of a noun phrase includes prepositions. See Figure 4.

2.4 Extracting Structural Relations

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, hooking up articles, adjectives and modifier nouns to their head nouns. Then, noun phrases are scanned right to left, connecting nouns over prepositions. Then, starting from verb phrases, phrases are scanned before the verb phrase for an unconnected head which becomes the subject, and likewise to the right of the verb for objects. Passive and active voices are treated, and some relative pronouns

```
plasma , maternal < ADJ
level , maternal < ADJ
plasma , fetal < ADJ
level , fetal < ADJ
level , plasma < NN
correlation , level < NNPREP
acid , glucose < NN
acid , free < ADJ
acid , fatty < ADJ
level , acid < NNPREP
...
```

Figure 5: Structural syntactic relations extracted. Ambiguous relations are maintained, for example 'maternal' may modify 'plasma' or 'level'. Both are retained.

are correctly handled, See (Grefenstette 1992).

Such a technique does not address any of the finer points of syntactic analysis, such as anaphora resolution, multi-word verbs, garden paths, etc. But is does handle a large percentage of natural language text, and is robust and rapid. Since we are interested in extremely large corpora, the information that is gleaned, though not complete is useful.

By similar considerations, when ambiguities arise, all options are retained. See Figure 5.

3 Similarity Calculation

At this point, we have extracted for each term a list of the words modifying it. If the term was modified by an adjective or by another noun, the frequency of that modification throughout the corpus provides one attribute for the term. If the term was in a relationship with a verb, that verb postfixed by an indicator of the relationship (-SUBJ, -DOBJ, -IOBJ) and the frequency of this relationship throughout the corpus form another attribute for the term. See Figures 6 and 7.

In order to calculate similarity, techniques developed in the social and natural sciences for classification purposes can be used. Each word, as shown in figure 6, can be considered an object and its collection of context features, attributes. Using methods, described for example in (Romesburg 1984) we calculate a similarity measure between every pair of words in

the text. We implemented a large number of similarity measures and found then the Tanimoto (1958) measure using log-entropy (Dumais 1990) weightings gave the best intuitive results. Each relation pair was given a local weighting of log(Frequency + 1) which was multiplied by a global weighting of the attribute involved, using

$$1 - \sum_j \frac{p_{ij} \log(p_{ij})}{\log(nbrels)}$$

where p_{ij} is

$$\frac{freq\ of\ attribute_j\ with\ object_i}{number\ of\ attributes\ for\ object_i}$$

and where *nbrels* is the total number of non-unique term-attribute relations extracted from the corpus. Our formula for the weighted Tanimoto similarity measure between two objects obj_m and obj_n, where the sums are over all unique attributes *att*, is

$$\frac{\sum_a tt \min(weight(obj_n, att), weight(obj_n, att)}{\sum_a tt \max(weight(obj_n, att), weight(obj_n, att)}$$

where the sums are over unique attributes. Note when the weights are restricted to 0 and 1 that this last formula is equivalent to a binary Tanimoto formula, though this is by no means the only way in which to generalize the binary formula to the weighted case.

This measure was used for the evaluation phase described below. Retaining the closest words to each word generates similarity lists as in Figure 8.

4 Evaluation

In order to test that these words were usefully similar, we took the queries that existed for commonly used testbed database of medical abstracts (MED), This database contained one million characters, and 160,000 words. Extracting the syntactic contexts of noun produced 71000 pairs of words. Of these, there were 53300 unique pairs. These pairs were composed of 5289 unique words that were compared

```
level absolute
level absolute
level accompany-DOBJ
level accompany-SUBJ
level accompany-SUBJ
level account-SUBJ
level achieve-DOBJ
level achieve-IOBJ
level achieve-SUBJ
level acid
level acid
...
```

Figure 6: Each word, here 'level', possesses contexts described by the words with which it enters into relations.

```
level=> concentration, value, excretion, content
```

'Level' and 'concentration' were close because they shared the following context words:

```
maternal fetal plasma level glucose
free fatty acid determine-DOBJ ffa
phospholipid rna rat observe-DOBJ
follow-SUBJ same low high alter-SUBJ
blood triglyceride tissue occur-SUBJ
nefa infant increase-DOBJ heparin value
produce-IOBJ serum raise-DOBJ increase
hour mean found different phosphorus
reduce-SUBJ reduce-DOBJ initial
metabolite relative remain-SUBJ
carbonyl mg selenium infect-SUBJ
depress-DOBJ phosphate amino-acid sugar
venous maintain-DOBJ maximum citrate
decrease-DOBJ calcium ca ml polyol
adipose-tissue umbilical sr
growth-hormone hgh gh
```

In this corpus, 'level' is close to 'excretion' because they share:

```
plasma level acid determine-DOBJ
follow-DOBJ rat low growth activity
alter-SUBJ study-DOBJ increase-DOBJ
dose begin-SUBJ normal woman increase
mean phosphorus corticosteroid steroid
patient reduce-DOBJ subject excretion
excrete-SUBJ urinary metabolite breast
advance-SUBJ estriol differ-SUBJ cancer
depress-DOBJ phosphate citrate
persist-SUBJ ca neutral spontaneous
fail-IOBJ ketosteroid bilirubin estrone
thallium
```

Figure 7: Each word is considered an object, with the words modifying it as its attributes. Similarity measures calculate the closeness of two objects using these attributes.

plasma => ffa, flow, blood, serum, excretion, glucose, level,

correlation => difference, reduction, rise, degree, pattern,

acid => concentration, content, level, activity, protein

ffa => insulin, sugar, glucose, utilization, calcium

glucose => ffa, serum, calcium, release, plasma, sugar

slope => order, start, prevalence, coefficient

line => culture, surface, bone-marrow, layer,

phospholipid => anatomy, lipide, pyruvate, stearate, polymerase,

change => increase, effect, response, study, pattern,

development => change, increase, response, incidence, growth,

course => severity, day, incidence, history,

day => hour, week, year, month, hr, time

rat => mouse, animal, dog, female, infant,

Figure 8: Some other words found similar in MED

using 9997 unique attributes. The closest terms to each term possessing at least 20 non-unique term-attribute pairs were calculated. 684 terms appeared this frequently. We iteratively expanded the queries by adding in the words closest to any of these 684 terms appearing in a query. By 'closest,' we accepted the word with the smallest distance (measured from 0 to 1.0) to the term, as well as any other word within 0.01 of this distance.

Queries were processed by using a standard cosine measure (Salton 1971) with log-entropy weighting recalculated using document occurrence of terms.

Keen to the the warning of Salton in relation to previous expansion studies (Salton 1972) about the sensitivity of the cosine measure to query length, in our system additional terms are considered true doubles of the original query terms and do not affect the query norm. Each occurence of an expansion term is treated as an occurence of the original term.

```
      Query  8:effect  drug  bone-marrow
                man  animal  pesticide
                significance  change
Expanded Query:effect  drug  bone-marrow
                man  animal  pesticide
                significance  change
                response  agent  marrow
                boy nucleoli  girl
                ulcerative-colitis  rat
                consideration  increase

      Query 21:language  development
                infancy  pre  school  age
Expanded Query:language  development
                infancy  pre  school  age
                speech change  increase
                education  range

      Query 23:infantile autism
Expanded Query:infantile autism psychosis

      Query 28:palliation  temporary
                improvement  cancer
                patient  drug
                x-ray  surgery
Expanded Query:palliation  temporary
                improvement  cancer
                patient  drug
                x-ray  surgery
                chemotherapy  shrinkage
                regression carcinoma
                case  agent  operation
```

Figure 9: Examples of query expansion using most similar words.

Examples of expanded queries, automatically generated by using words closest to the query terms can be seen in the figure 9.

The resulting expanded queries resulted in a slightly improved average precision of retrieval which is used as a standard measure in the Information Retrieval field (Salton 1971). The results are shown in the following graph.

The graph shows the average precision for all the queries(32) in the data set MED, at levels of 10% to 90%. A 20% recall level, for example, means going down the ranked list of documents, examining documents, until 20% of the relevant documents are found in the list. Before reaching all 20% of relevant documents for this query, some irrelevant documents may appear. The precision percentage at that recall

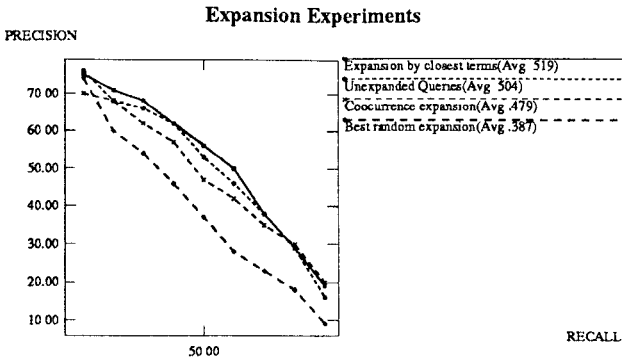

Figure 10: Average precision results on MED after query expansion

level tells what percentage of the documents examined were relevant.

This improvement was a pleasant confirmation of the intuitive feeling of similarity in the extracted wordlists. Some early success on query expansion using document co-occurence techniques had been reported (Sparck Jones 1971), though subsequent experimentation (Minker *et al.* 1972) (Smeaton and van Rijsbergen 1983) gave negative results.

To compare with these previous results, we ran experiments expanding the queries in two ways. In the first, we calculated the similarity between words using document co-occurrence data, as used in most of the previously published experiments.

This result is marked on the graph as "Cooccurrence expansion" and, as reported, gives a lower average precision at all levels of recall.

In a second experiment, we expanded queries by randomly adding words based on the added word's frequency in the corpus, a test also performed by (Smeaton and van Rijsbergen 1983). We ran this experiment 100 times with different random seeds, and results were always much lower than with using the unexpanded query. The best of these random test is plotted on the graph and labeled "Best random expansion."

The result of these experiments suggests that the words brought together as similar by our technique do indeed share close semantic ties to the original query terms for this database.

5 Related Research

Research related to the spirit of this technique can be found in (Hearst 1992) and (Ruge 1991). Hearst (1992) used lexico-syntactic patterns such as *NP* {, *NP* } *{,} *or other NP*

```
   Bruises, ..., broken bones or
other injuries
```
⇒ *hyponym("bruise", "injury")*
⇒ *hyponym("broken bone", "injury")*

to extract hyponymic relationships between words. These relations can then integrated into a hierarchical thesaurus, such as has been done for WordNet (Miller *et al.* 1990). As an evaluation of the relations found, the author showed that there was a good overlap between 106 relations that she extracted from *Grolier's American Academic Encyclopedia* using one such pattern, and a 34,000 word manually constructed WordNet hierarchy.

Ruge (1991) used a similar technique to ours, first extracting noun phrases from a corpus of 200,000 patent abstracts, and then calculating similarity of heads by comparing the words modifying them. Since each term was sometimes a head and sometimes a modifier, a similarity measure between two terms was developed that took into account the number of shared heads, when the terms were used as modifiers, and the number of shared modifiers when the terms were used as heads. She was able to find relations such as

```
container => enclosure, bottle,
   receptacle, cavity, vessel,
   tank, pouch
acceleration => deceleration,
   speed, velocity, inclination,
   movement, correction
```

efficient => economical, simple, effective, easy, compact, simultaneous, direct

This is very similar to our approach, though restricted to context within noun phrases only. As an evaluation of the results obtained, Ruge randomly chose 159 words from among the 8257 extracted and had a colleague select synonyms for each. Then a comparison of similarity measures was performed to see which brought the manually chosen synonyms closest to the top in the automatically generated similarity lists.

6 Conclusion

The future of Information Retrieval lies in knowledge-based techniques. We have presented here a technique and mentioned others which provide a portion of the needed knowledge automatically without using previous domain knowledge. Our test show that despite using

- imperfect morphological analysis
- imperfect syntactic disambiguation
- imperfect structural analysis
- limited contexts
- imperfectly understood similarity measures

we can nonetheless, over a large enough corpus, generate useful domain-specific semantic information. It can be hoped that improvement in any of the above items will improve and clarify the semantic information extracted.

References

(Crouch 1990) C. J. Crouch. An approach to the automatic construction of global thesauri. *Information Processing and Management*, 26(5):629–640, 1990.

(DeRose 1988) Steven J. DeRose. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39, Winter 1988.

(Dumais 1990) Susan T. Dumais. Enhancing performance in latent semantic (LSI) retrieval. *Unpublished manuscript*, 1990.

(Evans *et al.* 1991a) David A. Evans, K. Ginther-Webster, Mary Hart, R. G. Lefferts, and Ira A. Monarch. Automatic indexing using selective NLP and first-order thesauri. In *RIAO'91*, pages 624–643, Barcelona, April 2–5 1991. CID, Paris.

(Evans *et al.* 1991b) David A. Evans, Steve K. Henderson, Robert G. Lefferts, and Ira A. Monarch. A summary of the CLARIT project. Technical Report CMU-LCL-91-2, Laboratory for Computational Linguistics, Carnegie-Mellon University, November 1991.

(Grefenstette 1992) G. Grefenstette. Sextant: Extracting semantics from raw text implementation details. Technical Report CS92-05, University of Pittsburgh, Computer Science Dept., February 1992.

(Hearst 1992) Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*. COLING'92, Nantes, France, July 1992.

(Hindle 1989) D. Hindle. Acquiring disambiguation rules from text. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 118–125, Pittsburgh, 1989. ACL.

(Lewis and Croft 1990) D. D. Lewis and W. B. Croft. Term clustering of syntactic phrases. In J.L Vidick, editor, *13th International Conference on Research and Development in Information Retrieval*, pages 385–404, New York, September 5-7 1990. Association for Computing Machinery.

(Miller *et al.* 1990) George A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244, 1990.

(Minker *et al.* 1972) J. Minker, G. A. Wilson, and B. H. Zimmerman. Query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8:329–348, 1972.

(Peat and Willet 1991) Helen J. Peat and Peter Willet. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5):378–383, 1991.

(Phillips 1985) Martin Phillips. *Aspects of Text Structure: An investigation of the lexical organization of text.* Elsevier, Amsterdam, 1985.

(Romesburg 1984) H. C. Romesburg. *Cluster Analysis for Researchers.* Lifetime Learning Publications, Belmont, CA, 1984.

(Ruge 1991) Gerda Ruge. Experiments on linguistically based term associations. In *RIAO'91*, pages 528–545, Barcelona, April 2–5 1991. CID, Paris.

(Salton 1971) G. Salton. *The SMART Retrieval System: Experiments in automatic document processing.* Prentice–Hall, Englewood Cliffs, NJ, 1971.

(Salton 1972) G. Salton. Comment on "query expansion by the addition of clustered terms for a document retrieval system". *Information Storage and Retrieval*, 8:349, 1972.

(Smeaton and van Rijsbergen 1983) A. F. Smeaton and C. J. van Rijsbergen. The retrieval effectiveness of query expansion on a feedback document retrieval system. *Computer Journal*, 26:239–246, 1983.

(Sparck Jones 1971) Karen Sparck Jones. *Automatic Keyword Classification and Information Retrieval.* Butterworths, London, 1971.

(Sparck Jones 1991) Karen Sparck Jones. Notes and references on early automatic classification work. *SIGIR Forum*, 25(1):10–17, Spring 1991.

(Tanimoto 1958) T. T. Tanimoto. An elementary mathematical theory of classification. *I.B.M. Research*, 1958.

(Yu and Raghavan 1977) C. T. Yu and V. V. Raghavan. Single-pass method for determining the semantic relationships between terms. *JASIS*, 26(11):345–354, 1977.