Finding Semantic Similarity in Raw Text: the Deese Antonyms

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

Abstract

As more and more text becomes readily available in electronic form, much interest is being generated by finding ways of automatically extracting information from subsets of this text. While manual indexing and automatic keyword indexing are well known, both have drawbacks. Recent research on robust syntactic analysis and statistical correlations promises that some of the intuitive advantages of manual indexing can be retained in a fully automatic system. Here I present an experiment performed with my system SEXTANT which extracts semantically similar words from raw text. Using statistical methods combined with robust syntactic analysis, SEXTANT was able to find many of the intuitive pairings between semantically similar words studied by Deese [Deese, 1954].

Introduction

With the wide extension of network connections, cheap memory and powerful machines, more and more textual information is available on-line than ever before. Idiosyncratic collections of texts can freely and rapidly be amassed, and as quickly dispersed. Wading through such potentially ephemeral information in an efficient and profitable way poses a problem that has stimulated research.

The problems of size, diversity, and cost-effectiveness have led some researchers [Church and Hanks, 1990][Evans *et al.*, 1991a][Hearst, 1992] to examine what type of structure can be extracted from text using knowledge-poor and completely automatic approaches while not rejecting advances in robust linguistic parsing techniques.

I have been developing a domain-independent system for extracting similar words called SEXTANT. SEXTANT begins by performing sentence diagramming: connecting adjectives to nouns, nouns to nouns, subjects to verbs, and verbs to objects. The result of this diagramming over a large corpus provides context for each term in the corpus. SEXTANT then compares the contexts of each term, that is, what words two terms may or may not have in common as a means of estimating the similarity of the terms in that document collection. For example if two terms in a corpus are the only things described as "furry", "big", and "leaping" then, although those two things may never be mentioned in the same document, we might consider them similar.

In the experiment I present here, SEXTANT was able to capture some of the accepted semantic similarity of common English words by this simple knowledge-poor comparison method.

Overview Of SEXTANT

Raw text is processed by SEXTANT by using the morphological package developed for CLARIT [Evans *et al.*, 1991b]. This package performs simple morphological transformations and dictionary look-up. The dictio-

nary provides for each input word a list of grammatical categories with normalized forms for each category. The output of this morphological look-up is fed into a Hidden-Markov Model disambiguator implemented by Rob Lefferts and David Leberknight at the Laboratory for Computational Linguistics (CMU). The disambiguator calculates the most likely path through each sentence, using the frequencies of grammatical tag pairs from the Brown Corpus as well as the frequency of word-tag combinations for each word in that Corpus.

At this point, each word from the original text is tagged with a single grammatical category and normalized form. SEXTANT takes this tagged text and divides each sentence into complex noun phrases and verb phrases. Each sentence is then parsed, using ideas proposed in [Debili, 1982], and techniques described in [Grefenstette, 1992a]. The result of the parsing is a list of terms and the attributes modifying them. These term-attribute pairs are what SEXTANT uses to compute similarity between terms.

For example, the first sentence in this section produces the following term-attribute pairs in SEXTANT:

```
text raw
text process-DOBJ
using process-SUBJ
SEXTANT process-IOBJ
package use-SUBJ
package morphological
package develop-DOBJ
package use-DOBJ
package clarit
```

Notice that this parsing is incomplete, more complex relations such as `SEXTANT process-SUBJ` escape detection.

Comparison of term similarity is performed by SEXTANT using one of the distance-similarity measures developed in the psychology and social sciences [Romesburg, 1984](p.150). I have found that the Jaccard measure produces satisfactory results. The Jaccard measure calculates similarity by dividing the number shared attributes between two objects by the total number of attributes recognized for the two objects. This produces a value of 1 if both objects share all their attributes, down to a value of 0 if none are shared.

To temper the effect of common attributes, I developed a weighted Jaccard similarity measure, see [Grefenstette, 1992a] for details. The weights of each attribute are calculated using two weightings, a global weighting expressing the entropy of the attribute over the corpus (frequently appearing attributes have lower weights) , and a local weighting which is the log of the frequency of the term-attribute pair over the corpus. The weight for an term-attribute pair is the product of its global and local weightings, as was used in [Dumais, 1990].

Experimentation

Elsewhere [Grefenstette, 1992b][Grefenstette and Hearst, 1992] it has been shown that, for individual nouns, the contexts described above are sufficient for extracting semantically close words from a corpus, at least for nouns appearing sufficiently frequently in the corpus. Here I will describe an experiment comparing the modifiers, that is the modifiers become the objects to compare and the terms that they modify become the attributes.

Corpus Extraction

The experiment was done on a corpus extracted for some other experiment from a large online encyclopedia. The corpus deals principally with sports. It was created by extracting every sentence in the encyclopedia which contained one of the words from a list of one hundred sport terms:

`acrobatics`	`alai`	`angling`
`archery`	`association`	`athletic`
`athletics`	`badminton`	`baseball`
`basketball`	`battledore`	...

These terms are the hyponyms of "sport," taken from WordNet[Miller *et al.*, 1990]. The resulting corpus was 6 MB of text. The corpus

was parsed as described above, and SEXTANT compared the modifying words among themselves, using what they modified as attributes. Each modifier was compared with each of the 27403 other unique modifiers found in the diagrammed corpus.

Results

For each modifier SEXTANT provided the ten closest modifiers, using the Jaccard similarity measure of closeness and the terms that the words modified as the modifiers' attributes. During the evaluation of the results, I noticed that many of the modifiers found to be most similar were often near synonyms (as was already seen [Grefenstette, 1992b] for nouns). For example,

the closest words to	*was*
game	sport
major	important
different	various
film	stage
human	animal

But I also found that a great number of closest modifiers seemed to be antonyms.

In the early sixties, Deese[1954] studied human associativity of commonly occurring antonyms. Recent work [Justeson and Katz, 1991] has shown that a particular set of these antonyms, identified by Deese, appears much more often together in the same sentences than chance would dictate. I decided to extract the antonyms given in this set from SEXTANT's results. The closest words to each of the following Deese antonyms are given in the appendix:

active-passive	alive-dead	back-front
bad-good	*big-little	*black-white
bottom-top	clean-dirty	*cold-hot
dark-light	*deep-shallow	dry-wet
easy-hard	empty-full	*fast-slow
happy-sad	*hard-soft	*heavy-light
*high-low	*large-small	*left-right
*long-short	*narrow-wide	new-old
old-young	rich-poor	pretty-ugly
right-wrong	rough-smooth	short-tall
sour-sweet	*strong-weak	*thin-thick

A few other pairs were described by Deese (alone-together,far-near , few-many, first-last, single-married, inside-outside) are not present in the results since one or the other member was not tagged as a adjective/noun modifier.

In fourteen cases (marked with an asterick above) of thiry-three pairs, SEXTANT found that a word was closest or next-to-closest to its partner in the Deese list. Consider that this is compared to 27000 other candidates in the corpus. It truly seems that this technique combining low-level syntax and statistics has captured part of English intuition of word association.

Some words have bizarre associations, such as "tall-iconostasis" and "top-star". "Iconostasis"" only appears three times as a modifier, and modifies "altar, screen, clergy." "Tall" also modifies "altar" once. The only word in common between "short" and "tall", the Deese association pair is "leg", once in this corpus. But since "short" appears with 237 other attributes, "tall" and "short" are not seen as close by the Jaccard measure. This is a problem with the technique whenever a word does not have enough context with which to judge its similarity.

As for the pair "top-star", they share the following attributes : "center, command, grain, june, player(3 times), quarterback, side, star, surface(8 times), team(2 times), and union" whereas "top-bottom", the Deese pair share only "layer(3 times), line, side, surface(3 times), and water."

But in other cases, even though an exact match between Deese pairs are not made, the result is close. "Dry-moist", "poor-good", "front-hind", "smooth-resistant", "sweet-bitter" are semantically close to the corresponding Deese pairs. Words appearing most often ("large, small, long, high, light, black, white, low, strong") give the best results, as can be seen in the appendix where the words are presented in decreasing frequency.

Related Research And Conclusions

Over the Brown corpus, [Justeson and Katz, 1991] showed that antonyms tend to occur

more often together in the same sentence than chance would dictate. They were not able to say if these antonyms appeared together more frequently than any other adjective in that corpus. In SEXTANT, two words need not appear in the same sentence, or the same document, in order to be recognized as similar.

Ruge [1991] used an approach similar to SEXTANT, comparing modifiers and heads of noun phrases extracted from a large corpus of patents. She found that semantically similar words seemed to be modified by similar words, as well as to modify similar words. Hindle [1990] examined nouns which are subjects and objects of the same verbs, and produced similar results to that produced by SEXTANT. SEXTANT expands the contexts used to judge a word's similarity to interphrasal connections, providing for a richer context for each word.

These techniques of combining robust syntax with statistical comparison promise to be a rich and fertile area for semantic extraction. Working on any natural language corpus, they allow for a corpus-defined semantic extraction. These initial experiments bring back exciting results over large corpora, with no need for semantic modeling of the domain. More sophisticated statistical techniques, or richer word tags, may provide even better results.

References

Church, Kenneth Ward and Hanks, Patricia 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Debili, Fathi 1982. *Analyse Syntaxico-Semantique Fondee sur une Acquisition Automatique de Relations Lexicales-Semantiques.* Ph.D. Dissertation, University of Paris XI, France.

Deese, J. E. 1954. The associative structure of some common english adjectives. *Journal of Verbal Learning and Verbal Behavior* 3(5):347–357.

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

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

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

Grefenstette, G. and Hearst, M. 1992. A knowledge-poor method for refining automatically-discovered lexical relations: Combining weak techniques for stronger results. In *AAAI Workshop on Statistically-Based NLP Techniques.* Tenth National Conference on Artificial Intelligence.

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

Grefenstette, G. 1992b. Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of SIGIR'92*, Copenhagen, Denmark. ACM.

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

Hindle, D. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh. ACL. 268–275.

Justeson, John S. and Katz, Slava M. 1991. Co-occurrences of anonymous adjectives and their contexts. *Computational Linguistics* 17(1):1–19.

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

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

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

Closest Words to DEESE *Antonyms*

MODIFIER [FREQ]	groups of similar words (closest to farthest) groups with almost the same similarity are separated by |'s
large [837]	SMALL | important surface major | water great field single
small [746]	LARGE | surface major | field water new form important area
new [663]	major state important | different diverse | modern various
high [547]	LOW | different area level | time surface energy average ma
long [449]	SHORT | time range year | small surface single front large
light [305]	HEAVY energy different surface | radiation wind sun wave ga
low [250]	HIGH | temperature | increase average | total air ocean win
short [238]	LONG | prose numerous foot distance story | simple verse wa
strong [235]	WEAK | movement good different natural special party import
black [235]	WHITE | woman american national | major history successful
heavy [196]	LIGHT solid | gas product oxygen excessive industry tempera
white [190]	BLACK | red blue | brown small color | dark hair permanent
shallow [168]	coastal lake deep_water DEEP | temperate inland | pond stre
young [166]	man | black popular school numerous child help age famous h
right [165]	LEFT privilege | amendment freedom legal equal constitution
wide [160]	NARROW broad | worldwide | international matter external wi
hot [149]	COLD | warm | cooler pure continuous | liquid return corros
dry [138]	moist | cold liquid river continental layer warm nearby | d
little [134]	considerable BIG oxygen | national sufficient temperature w
deep [132]	SHALLOW | valley pacific | rock saline earth subtropical de
cold [132]	warm | HOT | coastal nearby | tropical continental moist su
thin [127]	outer transparent | THICK gas metal bony upper layer | exte
active [125]	principal system movement development period policy service
rich [106]	coastal mineral vast salt diversity animal complex warm org
good [106]	poor excellent | change strong production available polluti
hard [93]	SOFT solid | leathery rock rough mineral brick sedimentary
full [91]	increase knowledge shape | variable different congress pack
poor [89]	good agricultural adequate health | moist water_supply vill
bottom [88]	zone ocean shallow lake | floor cloud gill wildlife layer f
top [87]	star upper horizontal black valley non-hodgkin_lymphoma foo
front [85]	hind leg | horizontal room frame broad tail | facade foot c
dark [79]	bluish bright cooler horizontal sun red | planet brown visi
narrow [75]	WIDE | edge rectangular floor tube end | broad proper woode
slow [71]	FAST | pitch paranormal fluid stroke continuous rapid rate
smooth [66]	convex resistant concave plane porous fragmentation | ceram
big [60]	hardy LITTLE modest deer magnon | buffalo north&american ba
rough [59]	divergence | cut damp | pottery phase outer environment ext
soft [58]	HARD loose | face solid dead alkaline gray rubber edge laye
dead [56]	slab_sided adult | regulation festival unruffled staffordsh
weak [55]	frequency | STRONG electromagnetic | angular motion perpend
sweet [48]	anticlimactic bitter | dill peach cucumber | oat hardy kern
thick [44]	outermost outer THIN | genital intact brownish lighter glaz
left [44]	RIGHT hind | counterclockwise southern&hemisphere permit el
back [43]	conical tongue | throat sharp lizard layout | take foot alt
wet [40]	moist lime clay favorable towel smokestack | warm damp sand
fast [31]	unreliable SLOW imaginary harvest | wake competitive energe
clean [26]	conserve potable subsurface | cooler amputation etch the&vi
easy [21]	amphibole | concise | electron_beam perch catch persia opaq
tall [17]	iconostasis | chaste seville traveler courtyard the&winter
bad [17]	ugly | decree | oliver hugo alphabet | pauline chair doubt
empty [15]	vacant | cone | french&entente the&blue&angel superintenden
passive [9]	cautious | asphyxial | toad vihuela countertenor gruff volu
alive [7]	hypertensive heart&failure schizophrenic international_busi
dirty [6]	granary | caracalla sweat les&escaldes subsoil | relic | wa
happy [5]	the&imaginary&invalid flare | tehuacan saskatchewan&river p
sour [4]	precocious | operator | alexander&graham&bell | spartina sa
wrong [2]	mike&tyson geographer fiancee octave david&herbert&donald |
ugly [2]	thomas&rowlandson | bad | man
sad [1]	petit | vienna le | theater