THESAURUSES FOR NATURAL LANGUAGE PROCESSING

Adam Kilgarriff

Lexicography MasterClass Ltd. and
ITRI, University of Brighton
Brighton, England
adam@lexmasterclass.com

ABSTRACT

We argue that manual and automatic thesauruses are alternative resources for the same NLP tasks. This involves the radical step of interpreting manual thesauruses as classifications of words rather than word senses: the case for this is made. The range of roles for thesauruses within NLP is briefly presented and the WASPS thesaurus is introduced. Thesaurus evaluation is now becoming urgent. A range of evaluation strategies, all embedded within NLP tasks, is proposed.

Keywords: Thesaurus, corpus, NLP, word sense

1. INTRODUCTION

All manner of NLP (Natural Language Processing) tasks need a thesaurus. Wherever we suffer from sparse data, it is appealing to substitute the missing facts about a word with facts about the class of words to which it belongs. There is also a long tradition of using thesauruses[1] in information retrieval.

In this paper we first define and explicate what we understand a thesaurus to be. We then present the case for the importance of thesauruses for NLP. Next we briefly describe our thesaurus and how it was produced. Finally we discuss thesaurus evaluation.

2. MANUAL AND AUTOMATIC

A thesaurus is a resource that groups words according to similarity.

Thesauruses such as Roget and WordNet are produced manually, whereas others, as in pioneering work by Sparck Jones (1986) and more recent advances from Grefenstette (1994) and Lin (1998) are produced automatically from text corpora. One might consider the manually-produced ones to be semantic, since lexicographers put words in the same group according to their meaning, whereas the automatically produced ones are distributional, since the computer classifies them according to distribution. However there are both theoretical and practical arguments against viewing them as different sorts of objects.

The theoretical argument refers to Wittgenstein's “don't ask for the meaning, ask for the use” (1953). When invoking meaning as an organising principle, we are invoking a highly problematic concept about which philosophers have argued since Plato, and they show no signs of stopping now. It is not clear what it means to say words in the same thesaurus cluster have similar

[1] There is some debate about the plural of *thesaurus*.The opinion of lexicographers at both Oxford University Press and Macquarie is that it is inappropriate to assign latinate plurals to English words where a latinate plural is üünot well-established, and in the case of *thesaurus* it is not, so I adopt the standard English plural morpology

meanings: the logician’s response that synonyms are words that can always be exchanged *salve vertitate* –without affecting the truth value of the sentence– tells us nothing about word senses, or about circumstances where one word is more apt or accurate then another, and probably implies there are no, or very few, synonyms. Justeson and Katz (1991) demonstrate how one supposedly semantic relation, antonymy, key to the mental lexicon for adjectives (Miller 1998), ceases to be mysterious exactly when it is re-interpreted as a distributional relation. To understand or evaluate any thesaurus, we would do well to consider the distributional as well as the semantic evidence.

The practical argument is simply that semantic and distributional thesauruses are both tools we might use for the same purposes. If we wish to know what thesaurus is best for a given task, both kinds are candidates and should be compared.

Some thesauruses, usually manual ones, have hierarchical structure involving a number of layers. Others, usually the automatic ones, simply comprise groups of words (so may be viewed as one-level hierarchies). Hierarchical clustering algorithms may be applied to automatic thesauruses to make them multi-level (though this is hard to do well). The more-than-one-level hierarchies produced by algorithm will generally be simple hierarchies. The hierarchies produced manually are not --which leads us on to the vexed question of word senses.

2.1 Word senses

Authors of manual resources view the objects they are classifying as word senses, not words, whereas automatic ones simply classify words. In automatic thesauruses, words may or may not occur in more than one class, according to their distributional characteristics and the algorithm employed. Authors of thesauruses have generally aspired to assign each sense to exactly one class. Viewed as a classification system for word senses, Roget's is a simple hierarchy (Roget 1987).

However word senses are problematic objects. Identifying a word's senses is an analytic task for which there are very often no straightforward answers and no satisfactory criteria of correctness. Dictionaries disagree very often disagree, and thesauruses have a different perspective again on what should count as a word sense. The first priority for authors of thesauruses is to give coherent meaning-clusters, which results in quite different analyses to dictionaries, where the first priority is to give a coherent analysis of a word in its different senses (Kilgarriff and Yallop 2000).

From a practical point of view, if we wish to use a thesaurus for an NLP task, then, if we view the thesaurus as a classification of word senses, we have introduced a large measure of hard-to-resolve ambiguity to our task. We will probably have to undertake word sense disambiguation (WSD) before we can use the thesaurus and this will turn at least one fifth of our input stream into noise, since state of the art performance levels for WSD are below 80% (Edmonds and Kilgarriff 2002). This is a high price to pay for using a word sense based thesaurus.

For these reasons, we choose to consider thesauruses as classifications of words (which may have multiple meanings and may be multiply classified): not of word senses.

From this perspective, even though Roget may have considered his thesaurus a simple taxonomy of senses, we view it as a multiple-inheritance taxonomy of words.

3. SOME USES OF THESAURUSES

Tasks which could benefit from a high-quality thesaurus include parsing, anaphor resolution, establishing text coherence and word sense disambiguation.

3.1 Parsing

A thesaurus contains salient information for many parsing tasks including the very hard ones (for English and probably other languages) of conjunction scope and prepositional phrase (PP) attachment.

3.1.1 PP-attachment

Compare

eat fish with a fork

with

eat fish with bones

PP-attachment problems occur in a number of syntactic settings. This one, where the pattern is Verb-ObjectNP-PP, is a very common one: does PP modify ObjectNP or Verb? A simple strategy is to find counts in a large corpus: is there evidence for PP modifying Verb, or for PP modifying ObjectNP, and if there is evidence for both, for which is there more evidence? But it will often be the case that there is no evidence for either. In such cases, a thesaurus can help: we may not have evidence for <*eat, with, fork*> or <*fish, with, bone*> (we assume lemmatisation and a noun-phrase head-finder) but we are more likely to find evidence if we expand *eat*, *fish* and *bone* out to their thesaurus classes: perhaps we find <*munch, with, fork*> or <*eat, with, spoon*> or <*haddock with bone*>. We do not expect to find much evidence for <*eat, with, bone*> or <*fish, with, fork*> even when we have expanded to thesaural classes. (Clearly, a scoring system is required and this may need to be quite sophisticated.)

3.1.2 Conjunction scope

Compare

old boots and shoes

with

old boots and apples

It is a hard problem to determine whether the shoes are old, and whether the apples are old. It cannot be determined with confidence without more context. However one fact suggesting that the shoes are old while the apples are not is that *boot* and *shoe* are close in the thesaurus, and thesaurally close items are frequently found in conjunction, so *boots and shoes* is a likely syntactic unit.

3.2 Bridging anaphor resolution

Bridging anaphors are those where a later expression in a text refers to an entity mentioned earlier in the text, but rather than use a pronoun or similar, the author has used different content words. For example,

> Maria bought a beautiful apple. The fruit was red and crisp.

The fruit and the apple co-refer. The proximity of *fruit* and *apple* in a thesaurus can be used to help an algorithm establish that *the fruit* is a bridging anaphor referring back to the apple.

3.3 Text cohesion

For many practical and theoretical purposes, it is valuable to be able to break discourses into segments, where each segment coheres. A key aspect of its cohesion is that the topic is the same throughout a segment but changes at segment boundaries. Various methods have been proposed, some of which rely on the same word being repeated within, but not across, segments. Others use a thesaurus and use the premise that words within the same thesaurus classes will tend to occur within, but not across, segments.

3.4 Word Sense Disambiguation

Consider the ambiguous noun *pike* which can mean either a fish or a weapon, and the sentence within which we wish to disambiguate it

We caught a pike that afternoon.

Pike is not a common word so there is probably no evidence at our disposal for a direct connection between *catch* and *pike*. However there is likely to be some evidence connecting *catch* to one or more word which is thesaurally close to *pike* such as *roach, bream, carp, cod, mackerel, shark* or *fish*. Given a thesaurus, we can infer that the meaning of *pike* in this sentence is the fishy one.

3.5 Ontologies (a dangerous use)

The roles for thesauruses described above might be called language-internal. They are to support improved linguistic analyses of the text.

The alluring next step is to move from a linguistic analyses to a representation of what the string means.

This is the point at which the relevant academic discipline changes from NLP, or Computational Linguistics, to Artificial Intelligence (AI).

A central concern for AI is inference. To be intelligent, an agent must be able to infer more from a statement than is directly present in it. From the statement that Fido is a cat, the agent must be able to infer that Fido is alive. The reasoning required is that cats are animals, animals are alive, so Fido is alive. A crucial component is the hierarchical structure of the ontology, which tells us that cats are animals.

Ontologies look a little like hierarchical thesauruses. Both are hierarchies and both have nodes labeled with strings like *cat* and *animal*.

If a thesaurus could be treated as an ontology, this would be extremely useful for AI. It would mean the English sentence *Fido is a cat* could be mapped into a knowledge representation language with the word *cat* mapping directly to a node in the ontology, so we then have many inferences following from an English sentence. AI's greatest problem is the "knowledge acquisition bottleneck" – the difficulty of getting knowledge into the system. If we could start to automatically turn English sentences into knowledge items, which can contribute to ontology-building, AI will be delighted.

However it cannot be the word *cat* that maps directly to the ontology, as some cats are jazz musicians, and we do not wish to infer that they are furry. So, for AI purposes, it must be a sense of the word. AI would like to use a thesaurus as a link between language and ontology, but for that, the objects in the thesaurus need to be word senses, not words.

This use of a thesaurus is driven by AI's knowledge acquisition agenda. It is not linguistically motivated. It does not address the theoretical or practical problems implicit in a thesaurus of word senses. The allure is great, notably now with the semantic web beckoning, but that does not mean it will work. Linking in to ontologies is one reason for using thesauruses in NLP, but it is a dangerous one.

4 THE WASPS THESAURUS

The goal of the WASPS project was to explore the synergy between lexicography and WSD, developing technology to support a lexicographer so that they can simultaneously develop an accurate analysis of the behaviour and range of meaning of a word, and provide input for high-accuracy word sense disambiguation. The resulting system, the WASPbench, is described, and results reported, in Kilgarriff and Tugwell (2001) and elsewhere.[2] The central resource for the WASPbench, which is also the input to the thesaurus, is a database of grammatical relations holding between words in the British National Corpus (BNC): 100 million words of contemporary British English, of a wide range of genres.[3]

4.1 Grammatical relations database

The items central to our approach are triples such as <*object, catch, pike*>.[4] As well as *object*, the grammatical relations we use include *subject, and/or* (for conjuncts)*, head, modifier*; the full set is given in the reference above. To find the triples, we need to parse the corpus, which we do using a finite state parser operating over part-of-speech tags. The BNC has been part-of-speech-tagged by Lancaster University's CLAWS tagger, and we use these tags. The corpus was also lemmatized, using morph (Minnen et al 2000). In this way we identified 70 million instances of triples. For each instance, we retain a pointer into the corpus as this allows us to find associations between relations and to display examples.

The database contains many errors, originating from POS-tagging errors in the BNC, limitations of the pattern-matching grammar, or attachment ambiguities.

[2] Papers available at http://wasps.itri.brighton.ac.uk

[3] http://info.ox.ac.uk/bnc

[4] And also 4-tuples such as <*PP, eat, with, fork*>. Here we treat these as triples with the preposition or particle treated as part of the relation name, so this becomes <*PP_with, eat, fork*>.

However, as our interest is in high-salience patterns, given enough data, the signal stands out from the noise.

For language research purposes we present the information in the database on a particular word as a “word sketch”, a one-page summary of the word's grammatical and collocational behaviour. A set of 6000 word sketches was used in the production of the Macmillan English Dictionary for Advanced Learners (2002), with a team of thirty professional lexicographers using them every day, for every medium-to-high frequency noun, verb and adjective of English. The feedback we have received is that they are very useful, and change the way the lexicographer uses the corpus.

4.2 Similarity measure

For thesaurus building, the task is to calculate similarity between words on the basis of the grammatical relations they both share. We use the measure proposed in Lin (1998), as follows.

We break the task into three parts, one for nouns, one for verbs, one for adjectives. The core method is not suitable for identifying cross-part-of-speech similarities. The simplest way to proceed would be to count the number of triples that any two words share. Thus, the presence in the database of <*object, drink, beer*> and <*object, drink, wine*> scores one point for the similarity of *beer* and *wine*. The similarity score between any two words would then be the total number of shared triples.

This might produce useful results but fails to use the frequency information at our disposal. The pair of triples <<*object, repeal, law*>, <*object, repeal, statute*>> counts no more towards the similarity of *law* and *statute* than does the pair <<*object, take, law*>,<*object, take, statute*>> even though *repeal*, being a far more specific verb than *take*, provides more information. We have also failed to take account of how frequent the triples are. The simple measure would tend to give very high frequency words as nearest neighbours to most words.

In response, rather than scoring 1 for a shared triple, we assign a score which takes account of how much information each triple provides: the product of the mutual information of the first triple, and the mutual information of the second. It is this that we then sum over all the triples that two words share.

This is a moderately complex sum, and we potentially had to perform it as many as a 100 million times, to compute all the pairwise similarities. The process was optimised by reducing all mutual information figures to integers and logs so the multiplications then became integer addition. We then used a sampling approach rather than exhaustive computation of all similarities. We randomly selected several hundred words and, for each word of the same word class, identified how close it was to each of the random sample. We then only exhaustively calculated similarity for pairs of words near the same random-sample items.

4.3 Thesaurus description

For each word, we have retained as its “thesaurus entry” all the words with a similarity score above a threshold: generally between one hundred and five hundred near neighbours. Evidently, most words will occur in the entries for many other words, and we have not consolidated the data into groups. Polysemous words tend to have words in their entry corresponding to each of their meanings, and occur in the entries for the words with which they share each of their meanings.

Thesaurus entries have been generated for 17844 nouns, 4033 verbs and 7274 adjectives. Entries for a few words (showing the top 29 items) are presented below; the full listings can be inspected on the WASPS website.

nouns

doctor: nurse teacher solicitor practitioner lawyer officer surgeon engineer journalist consultant parent scientist expert physician farmer policeman official driver worker gp colleague professional servant accountant student manager politician staff specialist

exception: exemption limitation exclusion instance modification restriction recognition extension contrast addition refusal example clause indication definition error restraint reference objection consideration concession distinction variation occurrence anomaly offence jurisdiction implication analogy

pot:bowl pan jar container dish jug mug tin tub tray bag saucepan bottle basket bucket vase plate kettle teapot glass spoon soup box can cake tea packet pipe cup

zebra: giraffe buffalo hippopotamus rhinoceros gazelle antelope cheetah hippo leopard kangaroo crocodile deer rhino herbivore tortoise primate hyena camel scorpion macaque elephant mammoth alligator carnivore squirrel tiger newt chimpanzee monkey

verbs

measure: determine assess calculate decrease monitor increase evaluate reduce detect estimate indicate analyse exceed vary test observe define record reflect affect obtain generate predict enhance alter examine quantify relate adjust

meddle: verse tinker interfere enmesh tamper dabble intervene re-examine domicile disillusion partake dissatisfy molest skill engross adjudicate treble research recess cuff enlighten accede impound toil legislate wrestle outpace profit waive

irritate: amuse disgust alarm perturb puzzle horrify astonish infuriate startle please anger reassure disconcert embarrass shock unsettle disappoint bewilder frighten upset stun disturb outrage distract flatter frustrate surprise impress

boil: simmer heat cook fry bubble cool stir warm steam sizzle bake flavour spill soak roast taste pour dry wash chop melt freeze scald consume burn mix ferment scorch soften

adjectives

hypnotic: haunting piercing expressionless dreamy monotonous seductive meditative emotive comforting expressive mournful healing indistinct unforgettable unreadable harmonic prophetic steely sensuous soothing malevolent irresistible restful insidious expectant demonic incessant inhuman spooky

awkward: uncomfortable clumsy tricky uneasy painful embarrassing nasty tedious unpleasant miserable shy abrupt nervous inconvenient steep ugly horrible boring awful difficult unwelcome odd unnatural cheeky strange slow ridiculous unexpected messy

pink: purple yellow red blue white pale brown green grey coloured bright scarlet orange cream black crimson thick soft dark striped thin golden faded matching embroidered silver warm mauve damp

One striking observation relates to the rhythm of language. Long words tend to have long near neighbours and vice versa. Latinate words have Latinate neighbours and anglo-saxon ones, anglo-saxon neighbours: compare *exception* with *pot.* In general, the quality of the list only deteriorates when there are not enough instances of the word in the corpus, as in the case for *meddle*, with 131 corpus instances. (The similarity between *meddle* and *verse* rests on the expression *well versed in*. One can *meddle in* the same sorts of things one can be *well versed in*: art, politics and affairs of various kinds.) We intend to base future versions of the thesaurus on substantially larger corpora. The statistics we use tend to result in common words being classified as similar to common words, and rarer words to rarer words.

5. EVALUATION

While, naturally, we believe our thesaurus is very good, improving on Lin's because of the wider range of grammatical relations and the balance of the corpus, it is not obvious how to make the comparison scientifically.

Lin’s own evaluation compared against manual thesauruses, assuming that the manual ones are known to be correct so can act as a gold standard, analogous to the manually-annotated corpora used for evaluation of other NLP tasks. As sketched above, there are two problems here. Firstly, simple accuracy: for all those other NLP tasks, the gold standard corpus is only of use if it is reliable, as measured by replicability. We have little reason to believe that manually produced thesauruses have a high level of replicability. Entries for the same word in different thesauruses show only limited overlap.

Secondly manual thesauruses aim to classify word senses while automatic ones classify words. This is not quite as bad as it sounds, since, as argued above, both are most usefully viewed as classifications of words (in all their meanings) but certainly gives rise to some incompatibilities.

Most painfully, manual thesauruses contain no frequency information, so give no indication that *dog* is more frequent in its 'animal' than in its 'derogatory term for man' sense. NLP tools have no way (without a corpus and a great deal of error-prone additional work) of discovering the skew of the frequencies. Programs using them treat the two meanings as equal. This is not helpful, and is a drawback to using manual thesauruses for the tasks that NLP wants to use them for. If an automatic thesaurus algorithm, when applied to a large English corpus, succeeded in replicating WordNet or Roget, it would be a remarkable intellectual achievement but, if it came without frequency information, it would be of limited use for NLP.

We do of course sympathise with Lin and others in their attempts to use WordNet and Roget for evaluation and are aware they were not using them because they were ideal, but for lack of alternatives.

So let us consider possible alternatives. It is of greatest interest to evaluate a system or resource according to how well it performs a task which we really want it to perform, so let us revisit the four NLP tasks for thesauruses listed above:

- Parsing
    - prepositional phrase attachment
    - conjunction scope
- bridging anaphor resolution
- text cohesion
- word sense disambiguation

We believe all of these provide fertile prospects for thesaurus evaluation. For PP attachment, bridging anaphor resolution, and word sense disambiguation, publicly available evaluation corpora exist, and can be used to compare the performance of the same method in three variants: (1) with no thesaurus, (2) with thesaurus A, (3) with thesaurus B. We plan to build an evaluation corpus for conjunction scope, and we are currently exploring evaluation methods for text cohesion.

6. SUMMARY

First, we have considered manual and automatic thesauruses, arguing that they are alternative resources for the same task. This involves the radical step of interpreting manual thesauruses as classifications of words rather than word senses. This is at odds with their authors' presuppositions but, it is argued, it is necessary if they are to be useful to NLP. As long as a thesaurus is viewed as a classification of word senses, its theoretical basis will be unsound and WSD (introducing at least 20% errors) will be required *before* it can be used. A thesaurus based on words, not senses, is hard for AI to use, but that is an AI problem, not an NLP one.

The range of roles for thesauruses within NLP was briefly described.

The WASPS thesaurus was introduced, and examples of its entries given.

We believe that thesauruses will play an increasing role in NLP, and for that to happen, we must start evaluating them in the context of the NLP tasks where they have a role to play. A range of thesaurus evaluations was proposed.

References

Edmonds, P. and Kilgarriff, A. Editors and Introduction: *Journal of Natural Language Engineering 8 (4)*, Special issue on Evaluating Word Sense Disambiguation Systems. 2002.

Grefenstette, G. *Explorations in Automatic Thesaurus Discovery.* Kluwer 1994.

Justeson, J., S. and Katz, S.M. Co-occurrences of antonymous adjectives and their contexts.

Computational Linguistics, 17: 1-19. 1991.

Kilgarriff, A. and Tugwell, D. WASPbench: an MT lexicographer's workstation supporting state-of-the-art lexical disambiguation. Proc MT Summit, Spain, 2001: 187-190.

Kilgarriff A. and Yallop, C. What's in a thesaurus. Proc. 2nd LREC, Athens 2000: 1371-1379.

Lin, Dekang. Automatic retrieval; and clustering of similar words. COLING-ACL Montreal 1998: 768-774.

Macmillan English Dictionary for Advanced Learners. Edited by Michael Rundell. Macmillan 2002.

Miller, K. J. Modifiers in WordNet. In *WordNet: An Electronic Lexical Database.* Edited by Christiane Fellbaum. MIT Press 1998.

Minnen, G., Carroll, J. and Pearce, D. Robust, applied morphological generation. In Proc. Intnl Conf on NLP, Israel, 2000: 201-208.

Roget, Peter Mark. *Roget's Thesaurus.* Original edition 1852, Longman Edition edited by Betty Kirkpatrick, 1987.

Sparck Jones, Karen. *Synonymy and Semantic Classification.* Edinburgh University Press. 1986.

Wittgenstein, Ludwig. *Philosophical Investigations* Blackwell 1953.