Sense Tagging in Action
Combining Different Tests with Additive Weightings

Andrew Harley & Dominic Glennon
Cambridge Language Services Ltd
64 Baldock Street
Ware
Herts SG12 9DT
England
andrew@oakleaf.demon.co.uk

Abstract

This paper describes a working sense tagger, which attempts to automatically link each word in a text corpus to its corresponding sense in a machine-readable dictionary. It uses information automatically extracted from the MRD to find matches between the dictionary and the corpus sentences, and combines different types of information by simple additive scores with manually set weightings.

1. Introduction

This paper describes a working sense tagger, which attempts to automatically link each word in a text corpus to its corresponding sub-sense in the *Cambridge International Dictionary of English* (CIDE). Much research elsewhere has gone into the generation of probabilities from corpora and the extraction of textual information from printed dictionaries. Our research has had the distinct advantage of being done alongside a large lexicographic team, who have been developing further the database used for the creation of CIDE. It has thus been possible to have very useful computational data expertly coded by hand. We have been able to concentrate on defining the specification of this lexical resource, encoding it and then making use of it, rather than on trying to extract or refine the desired information automatically from existing corpora or printed dictionaries.

2. Methodology

The tagger, at present, works on one sentence at a time. Each word in the sentence has a certain number of possible senses. The tagger assigns a score (initially 0) to each possible sense of each word. A number of different tagging process could then adjust any of these scores, increasing them for a positive match (e.g. a collocation that indicates a particular sense), decreasing them for a negative match (e.g. capitalisation indicating a particular sense to be unlikely). At the end of all these processes, each sense of each word will have a particular score. For each word, the sense with the highest score is assumed to be the sense meant in the context.

Simple additive weightings are also commonly used in the evaluation of chess positions by computers, where for example, a pawn less could score -100 and an open file for a rook +15. It is thus possible for a number of positional factors to outweigh more concrete material factors.

It would be possible to use multiplicative probabilities rather than additive weightings. Chess programmers tend to prefer additive weightings because they are far simpler to program and also more efficient. There are more rigorous rules for combining probabilities, but it is not clear how much benefit this gives if the original probabilities are only rough estimates anyway. Probabilities can be derived from training corpora, but it is acknowledged that these can vary enormously from corpus to corpus, e.g. on grounds of register (Biber 1993). Such methods are far more appropriate for work in restricted contexts, where representative training corpora can be more easily derived.

3. Procedure

Besides some simple tests for suffixes (for unknown words), capitalisation, register and frequency, the main tagging processes are the following:

3.1 Multi-word unit tagger

The CIDE database contains detailed information on both single words and multi-word units. For a word pair X Y (e.g. *has been*), the tagger is thus able to produce possible scores for X and Y as separate words, and for X Y as a multi-word unit throughout each tagging process. If a multi-word unit is found, it is given an initial additional score (a headstart over the words treated separately) proportional to the number of words in the unit minus 1, but this can easily be cancelled out by other scores.

As a learner dictionary, CIDE contains much examples text. This examples text forms a convenient hand sense tagged corpus, though with only one word (the headword) sense tagged in each example. Much research has been devoted to using just collocation information for sense disambiguation, even using contexts of as much as 50 words (Gale, Church and Yarowsky, 1992). We instead choose to look more at the immediate context around a word, by dividing collocation match weightings by the distance between the pair of collocating words, expecting subject domain tagging (see section 3.2) to deal with more long-range effects.

3.2 Subject domain tagger

Each entry in CIDE has been subject coded. A subject domain for the sentence is created by looking at the subject codes of each likely (from the tests so far) sense of every word in the sentence, and at any document information available about the subject domain of the article, e.g. a sports page. Then the subject codes of each sense of each word are compared with the subject domain for the sentence and the number of matches noted. The subject codes are arranged in a hierarchy, so for example, *Christmas* and *Passover* would match at some levels, despite not having exactly the same subject code. Long sentences can distort the results, so the weightings awarded to subject domain matches are divided by the number of words in the sentence.

3.3 Part of speech tagger

Our part-of-speech tagger is based on a series of rules, listing valid 'transition pair' sequences of grammatical tags. These pairs can be given weightings but the emphasis of the approach is on the list of valid pairs rather than the weightings assigned to each pair. Thus most valid pairs are given a standard weighting of 0. Six special intermediate tags have been created to reduce the number of tag pairs that need to be listed and to add 'partial parsing' to the process. These are:

p[and p] around noun phrases acting as subjects (i.e. expecting to be followed by a verb)
p< and p> around noun phrases acting as objects
p(and p) around adverbial or prepositional phrases, or sub-clauses

Thus, for example, a determiner may only be preceded by p[or p< or a pre-determiner. The p(and p) are a particularly powerful feature which enable intermediate phrases to be ignored. The tagger does not check for p) followed by the next tag, but rather looks back to what came immediately before the preceding p(and then does the transition pair match on that. Atwell (1987) has termed these kind of brackets "hyperbrackets" and considers a very similar approach to that we are now adopting, choosing himself instead to add hyperbrackets to already tagged text to enhance it with parsing information, but thereby losing the benefit these hyperbrackets can assign to the part-of-speech tagging process itself. One example of the possible benefit is in trying to make the distinction between a preposition, which is generally followed by what we term an object noun phrase as it will not be followed by a verb, and a subordinating conjunction, which is generally followed by what we term a subject noun phrase as it will be followed by a verb.

For a valid transition pair between two tags, the score is simply calculated by adding the maximum score (from the other tagging processes) for a sense that can have each grammatical tag to the transition pair weighting (usually 0). There are also some special features to cope with more long-range effects (e.g. singular nouns being followed by the 3ps form of the present simple, conjunctions tending to co-ordinate the same grammatical tags). Thus, all valid sequences can be given a score by adding up the relevant transition pair scores.

Our method is more ambitious but intrinsically less efficient than Hidden Markov Model approaches, although certain restrictions are applied to reduce the number of sequences to a manageable size (e.g. a limit on the number of nested brackets). More time also needs to be spent on rule development.

3.4 Selectional preference pattern tagger

The selectional preference pattern tagger checks verb complementation and selectional preferences, and also adjective selectional preferences. Lexicographers have specifically attached CIDE grammar codes (which give verb complementation patterns) to selectional preference patterns using a restricted list of about 40 selectional classes for nouns. The tagger translates these grammar codes into sequences of grammatical tags and super-segmental tags representing the possible sequences that may follow the verb, and then integrates these with the selectional preference patterns.

It is these resulting patterns that the pattern tagger uses to test the syntactic and semantic veracity of the tag sequences produced by the part-of-speech tagger. If

Tagger event	Weighting
Part of speech 'transition pair' not found	rejected
Verb complementation pattern failure	-80[1]
Capitalisation failure	-60
Multi-word unit match	+50 times (words in unit - 1)
Frequency	0 to +50[2]
Selectional preference failure (for each argument)	-40[3]
Register failure	-30
Lexical collocate match	+30 per (distance between words)
Functional collocate match	+20 per (distance between words)
Illustrative[4] collocate match	+10 per (distance between words)
Subject domain match (for each level)	+30 per (words in sentence)

the argument pattern (subject and objects) fail to match a tag sequence, this is considered a verb complementation pattern failure. When an argument is encountered, the class specified in the selectional preference pattern is matched against the possible classes for the word. Selectional classes are hierarchical in structure like subject domain codes (see section 3.2), so allowance is made for near-matches. Adjective selectional preferences are matched in a similar but more simple way. Each adjective is coded with the possible class(es) of the nouns which it may modify. The adjective class is matched against the class of the noun which it modifies using much the same scoring system as for the verbs.

Selectional preference pattern matching has proved one of the most useful of all tests. A good example is the sentence:

The head asked the pupil a question.

Here, the CIDE database gives the possible selectional classes for *head* as body part, state, object, human or device; for *pupil* as human or body part; for *question* as communication or abstract.

The verb *asked* with two objects can only have the pattern *human asked human communication.* Thus, all the senses can be correctly assigned just by using selectional preferences.

3.5 Refinement

There are three main processes involved in refining the tagger's performance:

* Refining the lexicographic data, or indeed adding whole new categories of lexicographic data (e.g. selectional preference patterns).
* Writing new algorithms ("taggers").
* Analysing the interaction between different tests, and refining the weightings used for each.

A hand-tagged corpus is of course very useful for performing the third of these processes in a rigorous manner. The next stage of our research is to use the test corpus (section 4) as a training corpus to fine-tune the weightings. The main weightings currently in use, which may be of interest to other researchers trying to combine different tests, are shown in the table.

An example of how different taggers can interact is given by the following two sentences:

He was fired with enthusiasm by his boss.
He was fired by his boss with enthusiasm.

The DISMISS sense of *fired* matches with *boss* at 3 levels of subject domain coding, thus scoring 30*3/8 = 11 for both sentences.

The EXCITE sense of *fired* has *with* as a functional collocate and *enthusiasm* as an illustrative collocate in CIDE, and thus scores 20/1 + 10/2 = 25 for the first sentence and 20/4 + 10/5 = 7 for the second sentence.

Thus, assuming no other taggers intervene, the sense tagger will make the best possible assignment for these two, admittedly rather ambiguous, examples.

4. Results

To test the tagging, we compared the results against a previously hand sense tagged corpus of 4000 words.

[1] a successful match scores +10 per argument matched

[2] certain common senses, like the determiner use of *a*, were given scores up to +100

[3] or -10 for each level mismatch in the selectional preference hierarchy

[4] used in a CIDE example but not emboldened as lexicographically significant

Each of the 4000 words was manually assigned with just one sense tag and the tagging program likewise assigned precisely one sense tag to each word. The results are thus strictly determined by the number of matching taggings, with no ambiguous coding allowed. (These criteria are somewhat over-strict as in some cases more than one tag could be considered acceptable, e.g. where there are cross-references in the dictionary or where there is genuine ambiguity.) In calculating the results, prepositions were deliberately ignored because they have been heavily "split" in CIDE, far more so than in other dictionaries (Lazar 1996). Any attempt at distinguishing these senses would have to rely heavily on selectional preferences for prepositions, which are yet to be implemented within the tagging program.

At the sense (CIDE guideword) level, with an average 5 senses per word, the sense tagger was correct 78% of the time. At the sub-sense level, with an average 19 senses per word, the sense tagger was correct 73% of the time.

The part of speech tagging was also tested on the same texts to similarly strict criteria (i.e. no ambiguous coding allowed) and found to assign the correct part of speech 91% of the time. Three other part of speech taggers were run on the same texts for comparison. Two taggers developed from work done at Cambridge University under the ACQUILEX programme assigned 93% and 87% correctly, while the commercial *Prospero Parser* performed best, assigning 94% correctly.

5. Evaluation

These results clearly need to be improved dramatically before automatic sense tagging can prove practically useful. Nonetheless, these results, especially at sub-sense level, compare favourably with other research in the area.

Ng and Lee (1996) have found only 57% agreement when comparing the same texts tagged according to the same dictionary senses by different (human!) research groups. Cowie, Guthrie and Guthrie (1992) have reported 72% correct assignment at the LDOCE homograph level (and a much lower level for individual sense assignment). Wilks, Slator and Guthrie (1996) comment that 62% accuracy can be achieved at this level just by assigning the first (therefore most frequent) homograph in LDOCE. Furthermore, Wilks and Stevenson (1996) propose a method which should apparently achieve 92% accuracy to that same level just by using grammatical tags.

It must be noted however that the LDOCE homograph level is far more rough-grained than the CIDE guideword level, let alone the sub-sense level, and that Wilks and Stevenson's approach on its own would, by its very nature, not transfer down to more fine-grained distinctions. Other research, such as Yarowsky's into accent restoration in Spanish and French (1994), which reports accuracy levels of 90%-99%, is again at a more rough-grained level, in this case that of distinguished unaccented and accented word forms.

While the sense tagging results are fairly encouraging, the part of speech tagging results are at present relatively poor. It thus seems sensible, especially noting Wilks and Stevenson's analysis mentioned above, to first run a sentence through a traditional part of speech tagger before trying to disambiguate the senses. In theory, we would expect information such as subject domain and collocations to help part of speech tagging to be more accurate, however slightly, but we have not yet been able to demonstrate this in practice.

6. Acknowledgements

This work was supported by the DTI/SALT funded project Integrated Language Database, and built on work funded by the EC funded project ACQUILEX II and on background material from Cambridge University Press.

References

Atwell, E., 1987, Constituent-likelihood grammar, *The Computational Analysis of English*, Longman

Biber, D., 1993, Using Register-Diversified Corpora for General Language Studies, *Computational Linguistics 19:2*

Cowie, J., L.Guthrie & J.Guthrie, 1992, Lexical disambiguation using simulated annealing, *Proceedings of COLING-92*

Gale, W.A., K.W.Church & D.Yarowsky, 1992, Using Bilingual Materials to Develop Word Sense Disambiguation Methods

Lazar, K.A., 1996, Breaking New Ground, *The Even Yearbook 2*

Ng, H.T. & H.B.Lee, 1996, Integrating multiple knowledge sources to disambiguate word senses: An examplar-based approach, *ACL Proceedings*

Procter, P., 1995, (ed.) *Cambridge International Dictionary of English*, CUP

Wilks, Y.A., B.M.Slator & L.Guthrie, 1996, *Electric Words: Dictionaries, Computers and Meanings*, MIT Press

Y.A.Wilks & M.Stevenson, 1996, The Grammar of Sense: Is word-sense tagging much more than part-of-speech tagging?

Yarowsky, D., 1994, Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, *ACL Proceedings*