University of Brighton

ITRI-97-01

Sample the Lexicon

Adam Kilgarriff

March, 1997

This work was supported by the EPSRC under Grant K18931, SEAL.

Information Technology Research Institute Technical Report Series

ITRI, Univ. of Brighton, Lewes Road, Brighton BN2 4GJ, UK
TEL: +44 1273 642900 EMAIL: firstname.lastname@itri.brighton.ac.uk
FAX: +44 1273 642908 NET: http://www.itri.brighton.ac.uk

Sample the lexicon

1 Introduction

Lexical sense-tagging is not a well-understood task. Primary evidence is that people very often do not arrive at the same tag for the same occurrence-in-context: the agreement rate between word sense tags in SEMCOR (Miller et al., 1994) and the Singapore version of the same task is just 57% (Ng and Lee, 1996). When a task is not well-understood, it is wise to find out more about it before doing a lot of it. To find out more about it, it is necessary to look closely at it. There is too much data to look closely at everything. The approved scientific procedure, in such circumstances, is to take a sample.

We shall learn most if we use our knowledge of the domain to structure the sample. The domain can be looked at as a population of texts, or as a population of lemmas[1], each associated with a population of occurrences-in-context. The latter is more useful. Our interest in a tagged corpus is for what it tells us about lemmas, not for what it tells us about the texts which have been tagged. Human tagging effort will best be spent on closely investigating a sample of lemmas, and, for each, examining the kind of polysemy it exhibits, and the tagging issues its corpus instances raise.

Hence the title of the paper: sample the lexicon. A further exercise in lexical semantic tagging should not aim to tag all words (or all content words) in a corpus, as in SEMCOR. Rather, it should first sample the lexicon, and then tag a sample of those lemmas' corpus instances.

The paper includes a worked example of how an appropriate sample can be arrived at, and, in the Appendix, the sample.

2 Counter-arguments

The case against sampling is essentially that you provide no sense-tagged data for most words. At some point, this will be an issue. It will be an issue when a sense-tagged corpus is the source used by an NLP lexicon to determine some aspect of a word's behaviour, and the lexicon needs to be of more than an experimental size. That point is still a long way off.

A lesser counter-argument concerns the output: with sampling, it is a set of tagged contexts for each lemma in the sample. Without sampling (except at the text level), as in SEMCOR, the output is a document with all words[2] tagged. This has the appeal of being simpler to conceptualise, and substantially more compact.[3] This is not a major consideration.

Another counter-argument is that word sense selections are mutually constraining, so a text like SEMCOR where the context words are disambiguated is of far more value than one where they are not. But we do not know how often one sense-selection provides critical information for another in the same sentence. Sampling will provide a well-founded data set for pursuing the question.

[1] Eg. dictionary headwords or similar; I use the word 'lemma' in this paper to avoid the type-token ambiguity associated with 'word'.

[2] Or, as in SEMCOR, all open-class words.

[3] Since most tagged words will serve as part of the context of several other words. In the sampling approach, all contexts would be stored as separate items, with no overlap.

3 Word sense disambiguation research

'Sense-tagging' (eg. associating words with their word sense, as taken from a lexicon such as WordNet or LDOCE) is a task which has excited great interest in recent years, and for which resources now already exist.

The research task is to sense-tag automatically. Hand-tagged corpora potentially serve two purposes: data for statistical training, and a gold standard against which systems can be evaluated. Almost all reported word sense disambiguation (WSD) work treats a corpus as a set of lemmas, each having a set of contexts associated with them. Success is measured lemma by lemma. (Yarowsky, 1992; Yarowsky, 1995) presents tables where, for a small number of lemmas, a success rate is given. Various authors note, and the tables confirm, that success rates differ widely between lemmas. Usually, results are provided for less than twenty lemmas, selected for assorted reasons: sometimes because earlier work published a success rate for the lemma, so results can be compared; sometimes because the senses have different translations into another language. No attempts have been made to make them a representative sample.

An alternative approach was taken in (Ng and Lee, 1996) and (Cowie, Guthrie, and Guthrie, 1992). Ng and Lee attempted to disambiguate all occurrences of 191 "most frequently occurring and most ambiguous" nouns and verbs in a corpus. Cowie et al. aim to disambiguate all words but a small number of stoplist items. In both cases, the published papers do not provide success rates on a lemma-by-lemma basis.

In no existing work is there any classification scheme for lemmas, such that we can ask "for lemmas of type X, how does algorithm Y perform?" Results are either lemma-by-lemma, or for a very large class of lemmas. This is a central problem with all attempts to evaluate WSD research.

There is, of course, the associated problem of determining which senses for a lemma the WSD system should aim to disambiguate – broad homographs, as in (Yarowsky, 1995), or fine dictionary-type senses, as in (Ng and Lee, 1996) – and in either case, what is an appropriate source of the senses. A fuller investigation of this question requires a close examination of the sense distinctions made in a lexical resource, in relation to a particular NLP task. This, again, requires sampling the lexicon.

4 Contrast with part-of-speech tagging

The case for sense-tagging is usually developed by analogy to syntactic tagging and its successes. However there are severe contrasts. Syntactic tags such as NOUN, VERB, NP etc. are uncontentious, and the definitions of the categories have been refined by grammarians over the years. For sense-tagging, there are no such general categories. Each lemma has its own categories, and the authority for any set of these rests with a particular edition of a particular dictionary.

The point of using people in a large tagging exercise is to provide a set of correct taggings. If people frequently do not agree, the argument for a large-scale exercise collapses. For syntactic tagging, given good training and a detailed manual of good practice, a high degree of agreement is possible. For sense-tagging, (Fellbaum et al., 1996) make it clear that, within the SEMCOR team, there was often disagreement, and as cited above, there was disagreement between the Singapore and Princeton tags almost as often as not.[4] It

[4](Gale, Church, and Yarowsky, 1992) go to some lengths to construct an experiment where judges do

seems likely that some lemmas are easy for people to agree on while others are not. Again, before we can so much as say to what extent a correct set of taggings is a possibility, we need to sample the lexicon.

The overall goal of syntactic tagging concerns parsing. This gives a focus to any exercise in syntactic tagging: the purpose of the exercise (at least from an NLP perspective) is to provide classifications a parser can use. For sense-tagging, there is no single task to which it contributes. Motivations include lexicography, information retrieval and machine translation. Different sets of senses are salient for each task. (For translation, a different set of senses is required for each language pair.) This makes it harder to define the task, and argues for extra care in planning and piloting.

5 Methods: 'lexical' or 'textual'

The SEMCOR approach to tagging might be called 'textual'; human taggers work through the text, token by token. The meaning and themes of the text is foremost in the tagger's mind, and for each token to be tagged, a new set of sense-definitions is read. The approach taken in the joint OUP/Digital research project Hector (Atkins, 1993), in which a large quantity of tagging was undertaken, was, by contrast, 'lexical'. The taggers worked lemma by lemma, tagging all the corpus instances for the lemma one after the other. In this way, the meanings and sense-distinctions of the particular lemma were foremost in the tagger's mind. My own experience of tagging[5] was that the bulk of the intellectual labour went into the close reading of the dictionary definitions: only when they were fully and clearly understood could non-obvious tagging decisions be made. Taggers will make more accurate decisions faster if they work lexically rather than textually.

The lexical method also promotes the use of patterns. When a tagger notices a recurring pattern in the corpus lines for a lemma, they are usually able to infer that that pattern always signifies a particular sense. A good tagging methodology will promote the use of patterns, as was done in Hector.[6]

6 The Sampling Scheme

The ideal sampling scheme would classify lemmas according to the type of problems they pose to a disambiguation system, and would sample from those populations. However, the taxonomy of problem-types is not yet available (and is indeed a goal of the exercise). The appropriate method is iterative:

- sample the lexicon according to any criteria which seem salient, and for which information is readily available;
- study the sample;
- feed the results of the study back into a revised sampling scheme.

agree over 95% of the time, but the price is that judges were asked only to say whether a pair of corpus instances exhibited the same sense or not, not to say what the sense was. Also their experiment looked at only nine lemmas.

[5]Reference deleted for anonymity.

[6]Software for the automatic discovery of such patterns, and the semi-automatic assignment of patterns to senses, is currently under development.

The second step would involve looking only at dictionary definitions in some iterations, and looking also at corpus instances in others. Where there were SEMCOR taggings for a word, they would be a useful input.

Three straightforward yet salient features to use for sampling are word class (eg, N, V, ADJ), frequency (as identified from a large corpus) and degree of polysemy (obtained through counting the number of senses given in a lexical resource).

6.1 Worked example for a first-pass sample

The worked example covers nouns only. Frequency and degree of polysemy were each divided into four bands, giving a sampling scheme comprising 16 classes, or cells. The resources used were WordNet and the British National Corpus (BNC).

For each noun in WordNet, the BNC frequency was established by adding frequencies for singular and plural forms. Its level of polysemy in the Collins dictionary (as given in WordNet) was found, and it was then assigned to the appropriate cell. The first number in each cell of Table 1 is the number of noun lemmas assigned to that cell.[7]

I then summed the frequencies for the lemmas in each cell, to give the second number, which is the number of word-tokens in the BNC accounted for by the lemmas in the cell (in millions).[8]

To move from these figures to a sample of lemmas and of corpus instances to be tagged, we must decide (1) how many lemmas from each cell are to appear in the sample, and (2) how many corpus lines are to be tagged, for each lemma in the sample. The numbers were assigned by the author, in a pragmatic approach suited to a medium-sized research project, observing the following constraints.

The number of lemmas selected for the sample from each cell should increase with

- the number of lemmas in the cell;
- the frequency-band for the cell (as we are more interested in common words than rare ones).

Numbers of lemmas for each cell were allocated as multiples of five, with a minimum subsample size of ten, with a target number of 200 for the whole sample. No subsamples were allocated to cells which accounted for neither many lemmas nor many word-tokens.[9]

The number of corpus lines to be inspected per lemma in the sample, for each cell, should increase with

- the frequency-band for the cell, and
- the degree of polysemy for the cell

as both these factors may be expected to give rise to a more complex pattern of word use, requiring more data to be understood. These numbers were allocated by assigning 400

[7]There are various anomalies and errors in WordNet, the WordNet-based morphological exception-lists, the CLAWS POS-tagging of the BNC, and, in particular, in the WordNet version of Collins polysemy information, so the lemmas and numbers in each cell are to be regarded as approximate.

[8]As there are 100M words in the BNC, these are also percentages.

[9]It might seem unnecessary to take subsamples of monosemous lemmas, since there would not appear to be a sense selection task for them. However, most lemmas are at least occasionally used in non-standard ways, and including monosemous lemmas would provide an opportunity for determining the scale of this phenomenon for lemmas where the issue was not complicated by dictionary polysemy.

corpus occurrences per lemma to the most frequent, most polysemous lemmas (category AZ in the table and appendix), and reducing the figure by 40 for every step to a lower-frequency or lower-polysemy cell.

The last line of each cell of Table 1 presents, first, the proposed subsample size, for that cell, in lemmas; second, the number of corpus lines to be tagged per lemma; and third, the product of these two numbers. The sum of the products across the 16 cells is 52,800, the total number of corpus instances to be tagged under this scheme.

For each cell, a subsample of lemmas was randomly selected. The sample comprising these parts is presented in the Appendix.

7 Conclusion

Future work in sense-tagging should proceed by selecting lemmas from the lexicon, and then tagging a set of corpus instances for each. WSD programs could then be evaluated on the sample set of lemmas, and we would discover the different classes of problems for human, or automatic, disambiguation that different kinds of words (or words presenting different kinds of polysemy) introduced. Such a sample, with indicative numbers of corpus instances, has been presented.

Appendix: A sample of English nouns

For decoding category names see Table 1. Words for which the BNC did not provide as many corpus instances as indicated in the table have been excluded.

AY woman support city bank government member effect father moment relationship

AZ year back difference light community law way man mother court use sense group authority face

BW mouth american restaurant chest discussion employee spokesman manufacturer leadership awareness

BX opportunity context adult noise ball conservative cash while proportion bill membership gift expense republic penalty drink employment ratio knife championship category son shop corporation efficiency

BY cloud officer energy arrangement winter cheek engineer daughter code institution recovery minority works competition region introduction magazine examination phase chip ring bread move village mechanism

BZ pattern pair impression hole supply height flight truth key reader preparation standard heart representation metal

CW spreadsheet european dumping zoo snag chap consonant adequacy londoner rejection broadcaster hospice colliery layout plight

CX armament dice keeping contractor statistics deletion hurry referee porch loom leisure semantics prohibition granite thickness motherhood essential magnate innovation melon

CY tariff priest dive reservoir favour trumpet cry mortar slate fraction synthesis pet curfew distortion mail

	Frequency band				
Num Senses	Top 200	Next 1,000	Next 5,000	Remainder	TOTALS
0–1	AW 3; .10 0	BW 76; .40 10x240=2,400	CW 1365; .94 15x200=3,000	DW 10,141; .52 20x160=3,200	11,585; 1.96 45; 8,600
2–4	AX 30; .80 0	BX 335; 2.27 25x280=7,000	CX 2,436; 2.20 20x240=2,800	DX 5,077; .40 20x200=4,000	7,878; 5.67 65; 3,800
5–9	AY 77; 2.20 10x360=3,600	BY 419; 3.22 25x320=8,000	CY 1,043; 1.22 15x280=420	DY 471; .05 0	2,010; 6.67 50; 15,800
10+	AZ 90; 3.08 15x400=6,000	BZ 170; 1.26 15x360=5,400	CZ 156; .24 10x320=3,200	DZ 29; .00 0	445; 4.58 40; 14,600
TOTALS	200; 6.18 25; 9,600	1,000; 7.15 75; 22,800	5,000; 4.60 60; 13,200	15,718; .97 40; 7,200	21,918; 18.92 200; 52,800

Table 1: The first line in each cell names the cell. The two numbers in the second line are the number of lemmas in that cell and the number of word-tokens in the BNC accounted for by those lemmas in the BNC. The numbers in the third line are a proposal for how the sampling scheme should be developed. The first number is a proposal for the size of the subsample of lemmas to be selected from the cell. The second is a proposal for the number of token per lemma-in-the-sample to tag. The third number, the product, is the total number of tokens to be tagged for that cell. So, for the cell AZ (representing the most common, most polysemous words); there were 90 lemmas in the category, and these 90 lemmas accounted for 3.08 million words in the BNC. I am proposing that 15 of these 90 words are included in the sample of lemmas, and, for each of these 15 words, 400 corpus instances are tagged.

CZ fellow spread knot discharge bolt puff jump grip float stroke

DW sac kerb qualifying humanist animosity airframe mystic chum anemone dick rectum tenet marshall raisin priory prairie eec blazer operand smog

DX sister-in-law upturn deformation absentee chub buttock mousse kinsman sunrise vestige glint rye feud mercenary pauper tycoon miniature devotee junta backlog

References

Atkins, Sue. 1993. Tools for computer-aided lexicography: the Hector project. In *Papers in Computational Lexicography: COMPLEX '93*, Budapest.

Cowie, Jim, Joe Guthrie, and Louise Guthrie. 1992. Lexical disambiguation using simulated annealing. In *COLING 92*, pages 359–365, Nantes.

Fellbaum, Christiane, Joachim Grabowski, Shari Landes, and Andrea Baumann. 1996. Matching words to senses in WordNet: Naive *vs.* expert differentiation of senses. In

Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press, Cambridge, Mass. forthcoming.

Gale, William, Kenneth Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings, 30th ACL*, pages 249–156.

Miller, George A., Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proc. ARPA Human Language Technology Workshop*.

Ng, Hwee Tou and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *ACL Proceedings*, June.

Yarowsky, David. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *COLING 92*, Nantes.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivalling supervised methods. In *ACL 95*, pages 189–196, MIT.