How Dominant is the Commonest Sense of a Word?

Adam Kilgarriff

Lexicography MasterClass Ltd. and ITRI, University of Brighton, UK
Email: adam@lexmasterclass.com

Abstract. We present a mathematical model of word sense frequency distributions, and use word distributions to set parameters. The model implies that the expected dominance of the commonest sense rises with the number of corpus instances, and that, particularly for commoner words, highly uneven distributions are to be expected much more often than even ones. The model is compared with the limited evidence available from SEMCOR. The implications for WSD and its evaluation are discussed.

1 Introduction

Given a word with multiple senses, how might we expect the frequency of the commonest sense to relate to the frequency of the other senses? This topic is important for Word Sense Disambiguation (WSD): if the commonest sense is commonest by far, accounting for, say, 90% of the corpus instances for the word, it becomes hard for an 'intelligent' WSD program to perform better than a dumb one that just always chooses the commonest sense, as 90% is hard to beat.

These issues were first explored by Gale, Church and Yarowsky in 1992 [5], who identify what they call the 'lower bound' for respectable performance of a WSD system as the score that a dummy system achieves if it simply always chose the commonest sense. The score for this system will be the proportion of the data accounted for by the commonest sense, as in the question of our title. Looking at the very small dataset available to them, they found an average figure of 70%. Their work has remained since as a cloud sitting over WSD: the lower bound issue (and a further set of concerns regarding the upper bound) continue to cast a shadow over much WSD activity and its evaluation [9,7,4]. While resources – notably SEMCOR [12] – are now substantially larger and more systematic than they were in 1992, they are still too small to give a general answer to the title question, and it remains open. In the absence of adequate resources for answering the question directly, this paper aims to give a new perspective on the issue indirectly through providing a mathematical model, and using a word frequency distribution to model a word-sense one.

After presenting and evaluating the model, we continue the discussion above covering the difficulties that the lower bound issue has created for the SENSEVAL[1] exercises and the relation between commonest sense and domain.

[1] http://www.senseval.org

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2004, LNAI 2448, pp. 1–9, 2004.

2 The Model

2.1 Assumptions

Firstly, we note that words have a Zipfian or power-law distribution [16]. As a first approximation, the product of frequency and rank is constant.

Next we assume no special relationship between different meanings of the same word. We assume any instance of a word can be assigned to one and only one sense. We assume polysemy is accidental and random, and that an n-way polysemous word can be modeled just as a set of n independent senses. While these assumptions are patently untrue [11,6,8] and indeed polysemy (as distinct from homonymy) is defined as senses bearing relations to each other, the assumption allows us to set up a mathematical model which can in due course be evaluated.

Then, in the absence of much empirical evidence about the distribution of the whole population of words senses, we assume their distribution is as for words. We expect word senses to be power-law-distributed. We assume that the population of word senses will show no interesting distributional differences to the population of words. (There will just be rather more of them.) Again, we make no claim that the assumption is true: its role is to enable the modelling.

In the formal part of the paper, we ignore senses with frequency zero in a given corpus, so a word with three senses of which one has no occurrences is treated as a word with two senses.

2.2 Simple Zipfian Model

Now, consider a two-way polysemous word of frequency n with senses $s1$ and $s2$. What can we say about the relative frequencies of $s1$ and $s2$?

The frequencies of $s1$ and $s2$ sum to n: that is $f(s1) + f(s2) = n$. For any m from 1 to $n-1$, we can have

$$f(s1) = m \qquad\qquad f(s2) = n - m \tag{1}$$

Let us consider two cases: the case where $m = 1$ and the case where $m = (n+1)/2$. (We first address the case where n is odd, as there is a special case where n is even.) The question is, if $s1$ and $s2$ are any senses, what should our expectations be about the relative frequencies of the commoner and less common sense?

Let us call the complete population of word senses in a corpus q. A two-sense word can then be created by randomly selecting any one of these q items, and then randomly selecting another. There are $q(q-1)/2$ possible pairs, so the complete population of 2-sense-word possibilities has $q(q-1)/2$ members. We now investigate the subset of these $q(q-1)/2$ items where the joint frequency is n. First, we work out the size of the subset. Then, for each member of the subset, we calculate the proportion of n accounted for by the commoner sense; the mean value for this proportion is then the expected value for the proportion of instances of a two-way-ambiguous word of frequency n to be accounted for by the commoner sense.

If word senses follow Zipf's Law in its simplest form, the product of a sense's frequency and its rank is constant. For lower-frequency items, the number of items having frequency x will be $k/(x(x+1))$ where k is a constant. (For the derivation, see [2, pp. 13–17].)

If $m = 1$, there will be $k/(1 * (1 + 1)) = k/2$ possible word senses that $s1$ might be. In that case, $s2$ has frequency $n-1$, so there are $\frac{k}{n(n-1)}$ senses that $s2$ might be. If we look at all the possible combinations where $m = 1$, we have any of the $k/2$ $s1$s combined with any of the $k/n(n-1)$ $s2$s, so the total is the product of $k/2$ and $k/n(n-1)$, that is, $k^2/2n(n-1)$.

If $m = (n+1)/2$, by similar logic, $s1$ may be any of

$$\frac{k}{((n+1)/2)((n+1)/2)+1)} = \frac{4k}{(n+1)(n+3)} \tag{2}$$

senses and $s2$ may be any of

$$\frac{k}{((n-1)/2)((n-1)/2)+1)} = \frac{4k}{(n-1)(n+1)} \tag{3}$$

senses. The total number of possibilities for an n-frequency word where the most frequent sense accounts for $(n+1)/2$ of the instances is the product of these two:

$$\frac{16k^2}{(n+1)^2(n-1)(n+3)} \tag{4}$$

So what is the relative likelihood of s1 and s2 being as near as possible to equally frequent, as against the skewed case where all but one of the data instances belong to s1? It is the ratio of these two numbers: $k^2/2n(n-1)$ and $16k^2/((n+1)^2(n-1)(n+3))$ or $32n : (n+1)^2(n+3)$.

If we take a random sample of words with frequency 101 ($n = 101$), then, on these assumptions, the ratio is 3232 to $102^2 \times 104$ or approximately 1:335. A 100:1 split is 335 times as likely as a 51:50 split.

2.3 Models using Brown and BNC word frequencies

The simple Zipfian distribution does not model word frequencies accurately, so it may be objected that it is unlikely to accurately model word sense frequencies. Improvements on the simple Zipfian model have been explored at length in the literature. ([2] is a book-length discussion of this and related questions; see [13] for an early critique and the development of the generalised power-law model.) All higher-accuracy models are parameterised, so an actual word frequency distribution is required for parameter-setting. Once we select an actual word frequency distribution, we may as well use it more directly to model word sense frequencies, and there is no longer any need to use Zipfian assumptions.

In the remainder of the paper, we use word frequency distributions from the Brown corpus [10] and the British National Corpus (BNC) [3] as two models for word sense frequencies. The frequencies were smoothed to give a monotone decreasing function.[2] Table 1 presents actual and smoothed values for a sample of frequency classes for both Brown and BNC.

[2] The value of each data point was recomputed using a linear approximation based on n data points surrounding the data point being recomputed. n was set to one seventh (rounded down) of the frequency class, so for the frequency class of words occurring 63 times, the smoothed value for the number of words occurring 63 times was calculated as one ninth of the sum of the number of words occurring 59, 60, 61,..., 66 or 67 times. The parameter was set to seven as this was the lowest value that gave a monotone decreasing function.

Table 1. Brown and BNC counts for a range of frequencies, raw and smoothed

Frequency	Number of words having that frequency			
	Brown		*BNC*	
	Raw	Smoothed	Raw	Smoothed
1	16278	16278.00	486507	486507.00
2	6097	6097.00	123633	123633.00
3	3543	3543.00	58821	58821.00
4	2249	2249.00	36289	36289.00
⋮	⋮	⋮	⋮	⋮
50	43	43.13	742	700.75
51	47	41.86	688	679.45
⋮	⋮	⋮	⋮	⋮
100	10	11.03	262	244.37
⋮	⋮	⋮	⋮	⋮

To replicate the calculation of the relative likelihood of a 1:100 as against a 50:51 split using smoothed Brown data, we note that the number of possibilities for a 100:1 split is $16278 \times 11.03 = 179,546$ whereas the number of possibilities for a 50:51 split is $43.13 \times 41.86 = 1805$.

On this model, the highly skewed split is 99 times likelier.

2.4 Generalised model

We now generalize the maths to take into account all possible values of m from $n/2$ to $n-1$, and thereby arrive at a Maximum Likelihood Estimate of the proportion of the data accounted for by the commonest sense, for a two-sense word of frequency n.

For each value of m, the number of possibilities is

$$V(m) \times V(n-m) \tag{5}$$

where $V(m)$ is the number of items having frequency m, and can in principle be drawn from a theoretical or an empirical distribution.

To find the average, for each of these possibilities, we need to add on the commonest-sense proportion that this value of m implies: m/n. We also need to accumulate the total number of possibilities that give rise to an overall frequency of n for the word, as the denominator. Thus we have, where n is odd,

$$\frac{\Sigma_{m=(n+1)/2}^{m=n-1} V(m) \times V(n-m) \times m/n}{\Sigma_{m=(n+1)/2}^{m=n-1} V(m) \times V(n-m)} \tag{6}$$

Where n is even and $s1$ and $s2$ are equally frequent, we cannot take the square of $V(n/2)$ to give the number of possibilities as that would be double-counting: the number of pairs in a set of t items is $t(t-1)/2$, so the number of pairs here is

$$e(n) = \frac{V(n/2) \times V(n/2) - 1}{2} \tag{7}$$

For even n, the expected value for the proportion accounted for by the most common sense is

$$\frac{e(n)/2 + \Sigma_{m=(n+2)/2}^{m=n-1} V(m) \times V(n-m) \times m/n}{e(n) + \Sigma_{m=(n+2)/2}^{m=n-1} V(m) \times V(n-m)} \quad (8)$$

Using these formulae and the Brown and BNC distributions, we arrive at the MLE's for the percentage of instances accounted for by the commonest sense, for various values of n, as shown in Table 2.

The analysis is also extended to the 3-sense-word case and the 4-sense-word case.

Table 2. MLEs for most common sense using Brown and BNC models

n	2-sense words		3-sense words		4-sense words	
	Brown	BNC	Brown	BNC	Brown	BNC
10	80.58	83.21	57.34	58.90	40.00	40.00
25	85.64	88.94	70.96	74.21	56.70	58.16
50	89.99	92.33	79.16	81.82	67.15	69.05
100	92.09	94.62	83.45	87.03	73.62	77.06
200	94.40	96.17	88.06	90.71	80.25	83.08
500	97.98	97.63	95.51	94.19	91.93	89.10

3 Empirical Word Sense Frequency Distributions

The sense-tagged SEMCOR database provides limited empirical evidence of word sense frequency distributions.

There were 55 words with two senses occurring in SEMCOR, for which the word frequency was 10.[3] The average percentage accounted for by the commonest sense, in this dataset, is 73.64%.

There were 41 3-sense words with frequency 10, and the average of the proportions accounted for by the commoner sense, across those 41 items, was 64.63%.

In the 'class' rows in Table 3 we have gathered together words across a small range of frequencies in order to give better averages. We have done this in a way that has kept the average frequency, for the class, at the value (10, 25, 50, 100) that supports comparison with Brown and BNC figures from Table 2; BNC figures for the equivalent category are copied across.[4] This means that the frequency ranges are slightly variable: the 96 items in the 2-sense 25 class had frequencies between 20 and 31, whereas the 70 3-sense words in the 25-class had frequencies between 20 and 30.

While most of the BNC figures are higher, the two sets of figures both show the same tendency for the commonest-sense proportion to steadily decrease with the level of polysemy and to steadily increase with the frequency.

[3] The 'word' here is lemmatised, so is equivalent to a dictionary headword. It covers only one word class, eg, noun or verb, so *crash* (noun) and *crash* (verb) are treated as distinct items.

[4] The category "4–6 senses" is clearly not directly comparable with the 4-sense case from Table 2.

Table 3. Proportion for commonest sense from SEMCOR, with BNC figures for comparison. The # column gives the number of words in SEMCOR that the data is based on: there were 55 words in SEMCOR which had a total SEMCOR frequency of 10 and for which two WordNet senses had non-zero frequencies. The % column states the average proportion accounted for by the commonest sense, across these # items.

n	2-sense wds			3-sense wds			4-6-sense wds		
	#	%	bnc	#	%	bnc	#	%	bnc
10	55	73.64	83.21	41	64.63	58.90	23	44.35	40.00
25-class	96	79.79	88.94	70	68.10	74.21	103	53.41	58.16
50-class	45	83.08	92.33	59	72.35	81.82	88	54.93	69.05
100-class	16	79.41	94.62	24	77.76	87.03	20	73.04	77.06

4 Discussion

We do not have empirical figures for large values of n, owing to the size of SEMCOR, but the fit between SEMCOR and BNC figures leads us to believe that word frequencies and word sense frequencies have similar distributions and we expect the skew to become more pronounced for higher values of n, as in Table 2.

The highly skewed split is to be expected much more often than 'even' one. One possible reason for the theoretical figures being higher than the SEMCOR figures lies in the dictionary-writing process. Where a lexicographer is confronted with a large quantity of corpus data for a word, then, even if all of the examples are in the same area of meaning, it becomes tempting to allocate the word more column inches and more meanings.

Consider the words *generous* and *pike*. *Generous* is a common word with meanings ranging from generous people (who give lots of money) to generous helpings (large) to generous dispositions (inclinations to be kind and helpful). There are no sharp edges between the meanings, and they vary across a range. Given the frequency of the word, it seems appropriate to allocate more than one meaning, as do all of the range of dictionaries inspected.

Pike is less common (190 BNC occurrences, as against 1144) but it must be assigned distinct meanings for fish and weapon (and possibly also for Northern English hill, and turnpike, depending on dictionary size), however rare any of these meanings might be, since they cannot be assimilated as minor variants. *Pike*-style polysemy, with unassimilable meanings, is the kind that is modelled in this paper. Where there is *generous*-style ambiguity, one might expect less skewed distributions, since the lexicographer will only create a distinct sense for the 'generous disposition' reading if it is fairly common; if the lexicographer encounters only one or two instances, they will not. Polysemy and frequency are entangled.[5] We should not be surprised to find actual data less skewed than the model predicts, though we may also note that *generous*-style ambiguity is probably much less important for NLP system accuracy than *pike*-style ambiguity, and it is plausible that NLP-critical ambiguity is more skewed, and more like our model, than dictionary-based ambiguity as exemplified in SEMCOR.

[5] The nature of this entanglement is explored further in [7].

4.1 The more data you have, the more senses you find

It may seem surprising that the 'commonest proportion' varies with n, the frequency of the word. It may seem to suggest that the ratio between an individual word's senses varies as corpus size increases, but it does not. The proportion changes because, in additional corpus data, we find additional senses for words which previously were monosemous, or which change from being 2-sense words to 3-sense words, or 3 to 4. Intuitively, the proportion increases with n simply because the ratio between $n - 1$ and 1 increases with n, and, since there are so many singletons, this ratio dominates the statistic.

An early finding from corpus-based NLP was that the more data you look at, the more word types you find, without end [15]. This also applies to meanings. As lexicographers also discover, the more data we study, the more meanings we discover.

4.2 WSD evaluation

For the two SENSEVAL exercises, the title question has complicated the evaluation in several ways.

Is commonest-sense information available? The lower-bound system which always chooses the commonest sense can only be implemented if it is known what the commonest sense is. For a WSD system that does not know, the baseline is hard to beat: for the SENSEVAL-2 English lexical sample task, the highest-scoring system which did not have access to that information scored 40% against a commonest-sense of 48% [4, Table 3, p 285]. For most languages and text types, such a resource is not available (although the ordering of senses in dictionaries generally follows a lexicographer's perception of importance, which is correlated with frequency, so dictionaries can provide indirect evidence of the commonest sense). The SENSEVAL organisers responded to this situation by dividing the systems to be evaluated into two sets: those that used a training resource (which gave word sense frequency counts, amongst other things) and those that did not. Both tasks are important: the resource-rich one is relevant for high-salience applications like WSD for general English, and the resource-poor one because, in the general case, training resources are not available.

Lexical sample The two options for evaluation explored in the SENSEVAL exercises were the lexical-sample route and the all-words route.

For the lexical sample, first, a set of words is selected; then, a set of contextualised instances of each of these words is selected. Whoever makes these selections implicitly sets the commonest proportion. It is likely that they will make selections which are biased towards 'even' splits. A word with 100 test instances, of which 99 all had the same sense, would not seem a good choice of a word for SENSEVAL, whereas a word with a 50:50 split would seem an entirely suitable candidate, even though, as we have seen, the former is far likelier.[6]

[6] An attempt was made by the author to address this issue when he organized English SENSEVAL-1, by including in the lexical sample one word, *amazing*, for which the sense inventory offered only one sense. It turned out that all instances were assigned to this sense – there were no unassignable instances – and everyone except the author was rather puzzled as to why the word had been included in the dataset.

All-words In the all-words approach, where all the content words in a text, or set of sentences, are used for evaluation, we encounter a different problem. As budgets do not support enormous sense-tagging exercises, for most words, not very many instances will be tagged. For most core vocabulary items, n will be low. SEMCOR contains only 220 words with frequency greater than 50. So we will not encounter the problem of the commonest proportion averaging over 90%, simply because samples are so small.

In neither case does the evaluation scenario reflect the scenario of an NLP system in use, with large throughputs of text.

The paper implies that the case that WSD can outperform the baseline has not properly been made, with results from SENSEVAL being biased and not properly addressing questions of lexical sample selection, or scale.

4.3 Identifying the commonest sense in a domain/corpus

The model suggests that the baseline performs remarkably well. But the baseline needs to know what the commonest sense is. A system that concentrates on identifying the commonest sense could well outperform one that concentrates on disambiguation.

The observation has been widely used in commercial Machine Translation (MT). While WSD (called Lexical Disambiguation in the MT community) is a central problem for MT, it is sidestepped, by using different lexicons for different domains, more often than it is addressed. While this approach may have been born of pragmatism rather than theory, the model in this paper tends to support it. If an NLP application is operating within a domain, it is cleverer to customise the lexicon for the domain (thereby reducing ambiguity) than to attempt to resolve ambiguity.

Within the NLP WSD community, similar effects have been observed. Gale et al. [5] note in a footnote

> It is common to use very small contexts (e.g., 5-words) based on the observation that people seem to be able to disambiguate word-senses basd on very little context. We have taken a different approach. Since we have been able to find useful information out to 100 words (and measurable information out to 10,000 words), we feel we might as well make use of much larger contexts.

In looking at a very large window, they approximate an approach which identifies a domain. Recent work by McCarthy et al. [14] has taken this strategy a step further, exploring in detail how different word senses are commonest in different domains, and how NLP application performance can be improved by using this information.

The structure of the SENSEVAL exercise, for SENSEVALs 1, 2 and 3, has not allowed systems to take this approach. At most a few sentences of context have been provided for each test example. There has not been any possibility of using very large contexts, and the opportunities for finding, for example, sets of ducuments sharing a domain with the sample instance (as in topic vector methods [1]) have scarcely been possible.

It is a commonplace that the words (and senses) we use depend on the sorts of things we are talking about, and that different word senses apply in different domains. The model presented in this paper suggests that finding which sense of a word is commonest (in a given corpus or subcorpus or document set) may reap great rewards, and that future SENSEVALs should find a way of crediting systems that take this approach.

Acknowledgement

The author would like to thank Diana McCarthy for her careful reading of the draft.

References

1. Eneko Agirre and David Martinez. 2000. Exploring automatic word sense disambiguation with decision lists and the web. *Proc. COLING Workshop on Semantic Annotation and Intelligent Content*, Saarbrücken, Germany.
2. Harald Baayen. 2001. *Word Frequency Distributions*. Kluwer, Dordrecht.
3. Lou Burnard, 1995. *The BNC Reference Manual*. Oxford University Computing Service.
4. Philip Edmonds and Adam Kilgarriff. 2002. Guest editors, special issue on evaluating word sense disambiguation systems. *J. Natural Language Engineering*, 8(4).
5. William Gale, Kenneth Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. *Proc. 30th ACL*, pages 249–156.
6. Patrick Hanks. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics*, 1(1):75–98.
7. Adam Kilgarriff and Martha Palmer. 2000. Introduction, Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs. *Computers and the Humanities*, 34(1–2):1–13.
8. Adam Kilgarriff. 1997. 'I don't believe in word senses'. *Computers and the Humanities*, 31(2):91–113.
9. Adam Kilgarriff. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech and Language*, 12(4):453–472. Special Issue on Evaluation of Speech and Language Technology, ed. R. Gaizauskas.
10. H. Kučera and W. N. Francis. 1967. *Computational Analysis of Present-day English*. Brown University Press.
11. George Lakoff. 1987. *Women, Fire and Dangerous Things*. Univ. Chicago Press.
12. Shari Landes, Claudia Leacock, and Randee Tengi. 1998. Building semantic concordances. In Christiane Fellbaum, ed., *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.
13. Benoît Mandelbrot. 1954. Structure formelle des textes et communications: deux études. *Word*, 10:1–27.
14. Diana McCarthy, Rob Koeling, Julie Weeds and John Carroll. 2004. Finding predominant senses in untagged text. *Proc. 42nd ACL*, Barcelona.
15. Donald E. Walker and Robert A. Amsler. 1986. The use of machine-readable dictionaries in sublanguage analysis. In R. Grishman and R. Kittredge, eds, *Analysing Language in Restricted Domains*. Lawrence Erlbaum, Hillsdale, NJ.
16. G. K. Zipf. 1935. *The Psychobiology of Language*. Houghton Mifflin, Boston.