

Distributional Lexical Semantics for Stop Lists

Mr Neil Cooke
University of Surrey
n.cooke@surrey.ac.uk

Dr Lee Gillam
University of Surrey
l.gillam@surrey.ac.uk

In this paper, we consider the use of techniques that lead naturally towards using distributional lexical semantics for the automatic construction of corpora-specific stop word lists. We propose and evaluate a method for calculating stop words based on collocation, frequency information and comparisons of distributions within and across samples. This method is tested against the Enron email corpus and the MuchMore Springer Bilingual Corpus of medical abstracts. We identify some of the data cleansing challenges related to the Enron corpus, and particularly how these necessarily relate to the profile of a corpus. We further consider how we can and should investigate behaviours of subsamples of such a corpus to ascertain whether the lexical semantic techniques employed might be used to identify and classify variations in contextual use of keywords that may help towards content separation in “unclean” collections: the challenge here is the separation of keywords in the same or very similar contexts, that may be conceived as a “pragmatic difference”. Such work may also be applicable to initiatives in which the focus is on constructing (clean) corpora from the web, deriving knowledge resources from wikis, and finding key information within other textual social media.

Stop words, lexical distributional semantics, information retrieval.

1. INTRODUCTION

Distributional lexical semanticists are interested in deriving information about words and the concepts they refer to using, for example, statistical approaches over vector space models. Part of the applicable hypothesis relates to Firth's comment regarding a word and its company. This is extended to considerations of lexical semantics in which words sharing similar company may relate to one another either through some shared super-type, a synonymy relation, or some variation therein. Pustejovsky, Anick and Bergler (1993) explored the combined use of a representation for lexical knowledge, a generative lexicon, and computations of Mutual Information over known syntactic classes. The authors cite Hindle's drink-based example, how it is possible to create a metric for the substitutability of a set of words within certain contexts depending on the extent of shared contexts of those words. For distributional lexical semantics: 'tea' and 'coffee' may share a number of collocations, e.g. drink. The more that are shared, the more likely it becomes that these items are either synonymous or belong to the same class. When investigated more robustly, beyond simple co-occurrence, this becomes a comparison of collocation pattern. In our work, we are considering whether models of collocation distributions can be used to derive stop word lists from arbitrary text corpora. The principle we use is similar to the work cited above, though we are attempting in part to derive the set of contexts, rather than the words themselves, based on sets of words selected at random from the corpus. Other research into the construction of stop word lists makes use of Zipf's law (Watts, Porter, & Zhu. 2002, Lo, et.al 2005, Trumbach & Payne 2007), similarity tests against pre-classified document sets (Wilbur & Sirotkin 1992; Hiemstra & de Jong 2001; Sinka & Corne 2003, Lo, et.al 2005), and specific selections of sets of reference data (Blei, et al. 2003; Fumera, et al. 2006). Such approaches result in a variety of stop word lists, hence a variety of corpus linguistics research being undertaken with variations in repeatability of this research. In addition, it is assumed that such corpora are “clean”, largely since they have been manually constructed. The stop lists tend to have a limited relationship with the corpus at hand, and a limited mathematical basis (Hiemstra & de Jong 2001). Our work draws on analysis of neighbourhoods of collocation patterns (Smadja 1993; Levy & Bullinaria 2001) and investigates statistically specific (narrow) peaks of these patterns, in combination with other measures used in characterising distributions in corpora (Gillam and Ahmad 2005). We consider, specifically the problem of separation of data based on pragmatic differences: the same word may be used in apparently identical syntactic and semantic ways, but much wider consideration is needed to establish the pragmatic differences. We initially discuss corpus profiling and Zipfian distributions, with consideration for how to identify certain types of problems with distributions within corpora (section 2), then discuss the Enron corpus, and our prior work relating to the problems of information pollution (section 3), and duplication of text “objects”, which may skew results of any analysis (section 4). We then discuss stop list research and the approach for calculating stop word lists and Distributional Lexical Semantics and present automatic extraction and distributional similarity examples from the Enron corpus and the Much-more

corpus to demonstrate our approach (section 5). Finally we discuss the need for sub-sampling, and some considerations for future efforts (section 6).

2. CORPUS PROFILING AND ZIPFIAN DISTRIBUTIONS

The distribution of tokens in a text is considered as an important property in information theory. Through manual analysis of James Joyce’s Ulysses, George Kingsley Zipf posited a power-law function using the notion that the rank of a token, when ordered by frequency, is inversely proportional to its frequency (Zipf 1949). Researchers have since identified that rank multiplied by relative frequency produces a constant around 0.1 (Li 1992, Yavuz 1974); rank multiplied by frequency produces a constant that is a tenth the number of tokens in the corpus (e.g. for the Brown Corpus, Manning and Schutze 1999, pp26-27).

Corpora, however, do not appear to be so well behaved. Consideration of the British National Corpus (BNC), comprising 100,106,029 word-tokens with 669417 different word-types, demonstrates this: Zipf’s law suggests that the constant is around 10 million. The data, on the other hand, do not follow this [Table 1] and, indeed, only between ranks around 500 to 3500 does this approximate easily.

Rank (r)	Frequency (f)	f * r	% Error
1	6187927	6187927	-38.19%
2	2941790	5883580	-41.23%
3	2682878	8048634	-19.60%
4	2560346	10241384	+2.31%
5	2150885	10754425	+7.43%
.....			
3001	3417	10254417	+2.44%
3002	3414	10248828	+2.38%
.....			
4991	1746	8712540	-12.95%
4992	1746	8714286	-12.98%
....			
65521	26	1703546	-82.98%
65522	26	1703572	-82.98%

TABLE 1: Error values - Zipfian expectations

Indeed, the same is true for the Brown Corpus. However, our consideration of Zipf’s law is not for the purpose of suggesting corpora are Zipfian, or to suggest that Zipf-Mandelbrot provides a better fit, but to suggest that Zipf’s law provides a simple computable benchmark against which characteristics of other corpora may be contrasted: we consider that deviations (error values in contrast to Zipfian expectations) may provide a signature for the distribution of the corpus.

High frequency words behave in certain ways, with the top 100 most frequent variously accounting for a large proportion of the tokens in the corpus – around 40-50%. But it is also interesting to consider that about 53% of words in the BNC have a frequency of 1 – and since the number of types is rather less than 1/10 of the size of the corpus, Zipf’s law breaks for obvious reasons. For words of frequency 1, their relative rank in a frequency list is arbitrary but their proportions are interesting. Consider frequencies from 1 to 10 in 5 specialist text corpora with a predicted derived value derived from Zipf’s law that suggests the proportion of words occurring with a given frequency, f, is 1/f(f+1): 50% for f=1, 12% for f=2, and so on. Values are shown in Table 2.

Frequency	Automotive	Nuclear	Finance	Nanotech	BNC	Predicted
<i>Tokens</i>	350920	393993	685037	1012096	100106029	
<i>Types</i>	14252	14937	28793	26861	669417	
1	38.44%	40.55%	36.76%	43.28%	52.84%	50.00%
2	14.14%	15.18%	15.20%	13.50%	12.92%	16.70%
3	7.14%	6.73%	8.58%	7.01%	6.02%	8.30%
4	5.10%	4.85%	5.78%	4.66%	3.71%	5.00%
5	3.64%	3.29%	3.99%	3.46%	2.54%	3.30%
6	2.69%	3.05%	2.96%	2.64%	1.90%	2.40%
7	2.15%	2.19%	2.30%	1.95%	1.51%	1.80%
8	1.78%	1.83%	1.82%	1.73%	1.20%	1.40%
9	1.80%	1.41%	1.54%	1.33%	0.98%	1.10%
10	1.48%	1.33%	1.45%	1.19%	0.83%	0.90%
Total	78.36%	80.42%	80.38%	80.76%	84.45%	90.90%

TABLE 2: Specialist Corpora, Error values - Zipfian expectations

Specialist corpora, it appears, are less prone to large vocabularies in comparison to the representative general language corpus (BNC) and the Zipfian prediction, which BNC most closely follows. Another consideration here is that these low frequencies may also reflect errors in spelling, and other untidy formulations within the source data.

On the other hand, it becomes a possibility requiring consideration that increased repetition is an artefact of repeated texts or, potentially, text fragments across numerous documents. We consider some of these issues in relation to the Enron corpus in the following section.

3. ENRON CORPUS “BURYING THE BAD NEWS”?

The Enron corpus is one of the largest business email collections available for public use from the Federal Energy Regulatory Commission (FERC). The most widely available version¹ comprises around 150 email accounts with 3349 folders, totalling some 517,431 files. Unlike most corpora, the Enron corpus was not designed or, originally, selected. It was largely a dump of the contents of the email server and hence presents particular challenges in addition to its size. Much Enron corpus research to date has focused on issues of data cleansing, for example, de-duplication, but also on elements of social network analysis, deception analysis and small scale automatic classification efforts. Table 3, published previously in, (Cooke, Gillam & Kondoz 2007a&b) provides a brief summary of this research.

Application	Corpus size (# emails)	Further Description
Automatic classification	12,500	Determining whether emails are for “Business” or “Personal” uses, University of Sheffield, UK, (Jabbari et al 2006)
Data cleansing, Preliminary analysis	200,399	Analysis of email threads and message distribution. Some folders removed, Carnegie Mellon (Klimt & Yang 2004)
Annotation; visualisation	1,700	Manual annotation of email categories. http://bailando.sims.berkeley.edu/enron_email.html
	255,636	Visualisation and clustering. Use of database structure separating bodies, headers and other elements. University of California, Berkeley http://bailando.sims.berkeley.edu/enron_email.html
Automatic classification	20,581	Automatic approach to building email folders Massachusetts Amherst (Bekkerman, McCallum & Huang 2004)
Data de-duplication	250,485	MD5 Hashes on body text to identify duplicates, resulting in 250,485 emails. Massachusetts, Amherst http://citr.cs.umass.edu/~corrada/enron/

TABLE 3: Related Enron corpus research

We initially consider two characterisations of the Enron corpus to explore how certain expectations may or may not hold: by frequency and by weirdness. In this analysis we contrast two versions of the corpus, the Enron (raw) 500k email Enron corpus¹ and Enron-CleanUCB a de-duplicated 200k email Enron corpus provided by Klimt & Yang (2004). We know in advance that the raw Enron email corpus provides at least two challenges, (i) vast amounts of duplication which distorts frequency analysis; (ii) presence of email headers, which distorts relationship analysis. Enron (clean) has, reportedly, removed email headers and duplicates.

The first 100 most frequent “words” in Enron-Raw account for 45% of the corpus; at frequencies up to 10, 77% of the vocabulary is accounted for with only 27% at a frequency of 1. We might summarise that such a corpus is demonstrating specialist tendencies, but this should be considered inconclusive. Contrasting the top 10 most frequent words between Enron-Raw and Enron-Clean (Table 4), it would appear that the initial list would have been dominated by elements of email headers (e.g. enron, com, x, cn); stop words have been impacted both by removal of duplicates and subject lines.

Enron-Raw		Enron-CleanUCB	
Word	Frequency	Word	Frequency
enron	7,555,888	the	2,806,643
com	6,881,814	to	2,025,907
the	5,684,275	r	1,896,214
to	5,072,137	and	1,286,641
x	3,654,791	of	1,180,840
and	2,593,183	a	1,077,861
of	2,391,399	in	862,776
cn	2,332,235	enron	766,304

TABLE 4: Enron Raw- Enron clean

It would appear, then, that the clean version would now be a much more appropriate corpus for further analysis. Consider, however, a tag cloud of tokens discovered with highest weirdness (highly frequent and unusual in comparison to the BNC) for both corpora (Figure 1 and Figure 2):

¹ <http://www.cs.cmu.edu/~enron/>

Collocate	Total	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
and	30584	623	300	7512	1961	3310	15262	768	230	372	246
may	15264	307	1144	161	10290	5	74	2815	249	190	29
contain	11004	1613	205	428	7	8630	0	0	38	83	0
the	13886	459	1485	681	160	473	151	551	966	667	8293
for	10129	493	377	149	88	116	105	296	629	7228	648
privileged	19390	68	13	1375	4780	1647	71	6599	2593	1398	846
material	6367	2	0	30	0	0	21	22	5122	1153	17
relevant	4863	4856	7	0	0	0	0	0	0	0	0
information	11143	688	1013	704	271	123	5379	338	1111	715	801
affiliate	5051	185	4855	7	0	0	0	0	0	0	4

TABLE 5: "Confidential" collocations BNC

Word	Total	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
contain	11004	1613	205	428	7	8630	0	0	38	83	0
privileged	19390	68	13	1375	4780	1647	71	6599	2593	1398	846
affiliate	5051	185	4855	7	0	0	0	0	0	0	4
legally	4724	3	1117	342	71	0	70	139	2475	499	8
intended	3990	69	17	4	0	0	10	2535	516	570	269
exempt	2480	0	0	0	0	0	0	430	218	1832	0
proprietary	3097	147	107	1726	70	7	649	258	129	3	1
unauthorized	1415	0	0	0	0	0	0	0	8	0	1407
solely	1399	0	0	1	0	0	0	9	1275	98	16
email	2864	63	510	932	506	1	5	4	21	5	817

TABLE 6: "Confidential" collocations BNC top 2000 removed

These collocates demonstrate strong distributional associations within this neighbourhood, and within the wider neighbourhood of our analysis, demonstrating the prevalence of confidentiality banners. In order to find any confidential material within such a corpus, it would first be necessary to subtract the confidentiality banners.

5. STOP! WORDS!

Previous analysis has demonstrated that there are several kinds of polluting information within the Enron corpus. Since these are artefacts of the corpus, we reflect on how we may remove stop words using a completely arbitrary stop list that has no relationship to the texts, nor any mathematical basis (Hiemstra & de Jong 2001). These distributions of pollution will cause difficulties to the uninitiated. Similarly, inappropriate stop lists can have severe effects on vector based semantic representations (Bullinaria & Levy 2006).

Typical approaches to constructing stop lists are to use Zipf's law (Watts, Porter, & Zhu. 2002, Lo et al 2005, Trumbach & Payne 2007) or comparison to other pre-classified document sets and performing a "similarity test" (Wilbur & Sirotkin 1992; Hiemstra & de Jong 2001; Sinka & Corne 2003, Lo et al, 2005), both of which have little foundation on the text under analysis.

Collocation analysis reveals many interesting shapes and patterns. We claim that collocations whose patterns look like flat lines across the collocation positions, up to five words to the left (L5 to L1) and up to five words to the right (R1 to R5) of the example nucleating word "energy" as shown in figure 4, have little relationship to the nucleate and hence have stop word characteristics. Typically, analysis methods use frequency as a means for stop list construction and standard deviation for ranking analysis, but ranking methods based on standard deviation are strongly biased by frequency. We are interested in words that have a strong distance relationship represented by a statistically significant narrow peak² and we claim that these have entropy and represent local context.

Our formulae below exploit both unusual frequency (*cohesion - strength of the peak*) and the narrowness of the peak (*local context - strength of the association*) to automatically rank the collocates and hence derive the stop list. We consider single peak patterns or "High Q" shapes, e.g. mission; flat lines, e.g. as, and locking exclusions e.g. *jones, vice, would*, as collocations with "energy" (Figure 4).

The High Q terms are combined with relative frequency using formula 1, 2 and 3. Relative Frequency is based on the whole corpus, we analysis the shapes that emerge from consideration of such patterns.

$$Q_peak = \frac{\sum_{i=1}^n F_i^2}{F^2} \quad (1)$$

F = total Frequency of the collocate

F_i = Frequency of the collocate at word distance $i = n$

n = number of collocation positions

²A high resonant Q in RF filter terminology

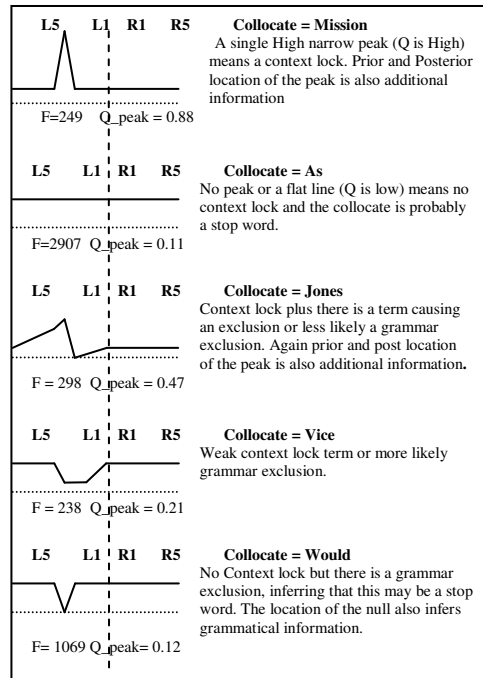


FIGURE 4: Dasovich³ "energy" collocation pattern examples

$$W = \frac{F_{SL} N_{GL}}{(F_{GL} + 1) N_{SL}} \quad (2)$$

F_{SL} = word frequency in target corpus
N_{GL} = reference corpus size (tokens)
N_{SL} = target corpus size (tokens)
F_{GL} = words frequency in reference corpus

Weirdness is exponential and combined as a cubed root to reduce its aggressiveness for unseen words.

$$Q_w = Q_{peak} \times W^{1/3} \quad (3)$$

To generate the stop list we arbitrarily chose 1.2 as the cut off point for Q_w , setting the value higher increases the Stop list aggression.

When reading, we use both knowledge (through relatedness) and observational experience (through similarity) to disambiguate words. Knowledge disambiguates through conceptual relationships: this approach requires both a validated knowledge base and heuristics (Budanitsky & Hirst, 2006). Observational experience uses Semantic Similarity and is based on human perceptions of frequently co-occurring words: if we observe that cat is frequently near animal then if platypus is also frequently near animal we may infer similarity between similarity between platypus and cat (Levy & Bullinaria, 2001). This is a lexical semantic similarity approach, and we can use such relationships to distinguish and associate. If two nucleates share strong High Q words then we can infer that there is a similarity, figure 5.

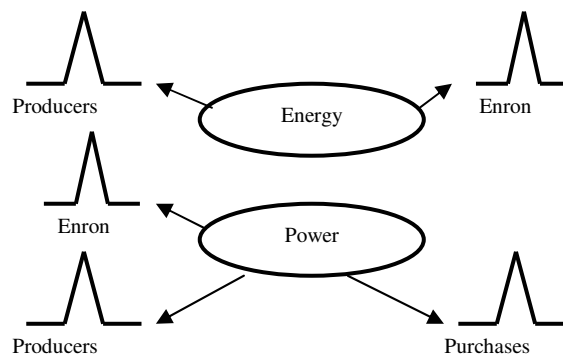


FIGURE 5: Similarity using Strength of association

³ Dasovich Email account consists of 21,433,538 tokens, 93,097 types, the Nucleate "energy" occurs 62,153 times, and the +/-5 token width vocabulary is 16821.

Levy & Bullinaria, use vectors based on collocate frequencies; we prefer a strength of association measure based on how unusually frequent and how tightly the distance relationship is constrained, hence Qw. We derive the formula 4 to compare similarities:

$$S = \frac{2 \sum Qw(a)_i \times Qw(b)_i}{\sum Qw(a)_i^2 + \sum Qw(b)_i^2} \quad (4)$$

S = Similarity where 1 = Identical
i = matching collocate

$Qw(a)_i$ = Qw of collocate *i* of nucleate *a*

$Qw(b)_i$ = Qw of collocate *i* of nucleate *b*

When the comparison is made between identical corpora, the weirdness measures will cancel out. This allows us to compare variations in term use across different subsamples i.e. examine subsample variations.

5.1 Jeff Dasovich’s energy, power & contracts

We initially removed all Dasovich3 energy collocations of frequency less than 11, since our confidence in the peak analysis results across 10 collocation positions would be low. This left 3992 collocations to analyse.

We ranked the collocations of energy the formulas for Q_peak, Weirdness and Qw, using a collocation window size of +/-5, and then compared the rankings against the BNC top 2000 stop list. Figure 6 below shows how these matched BNC stop list words are distributed (as a % accumulation) within these different rankings.

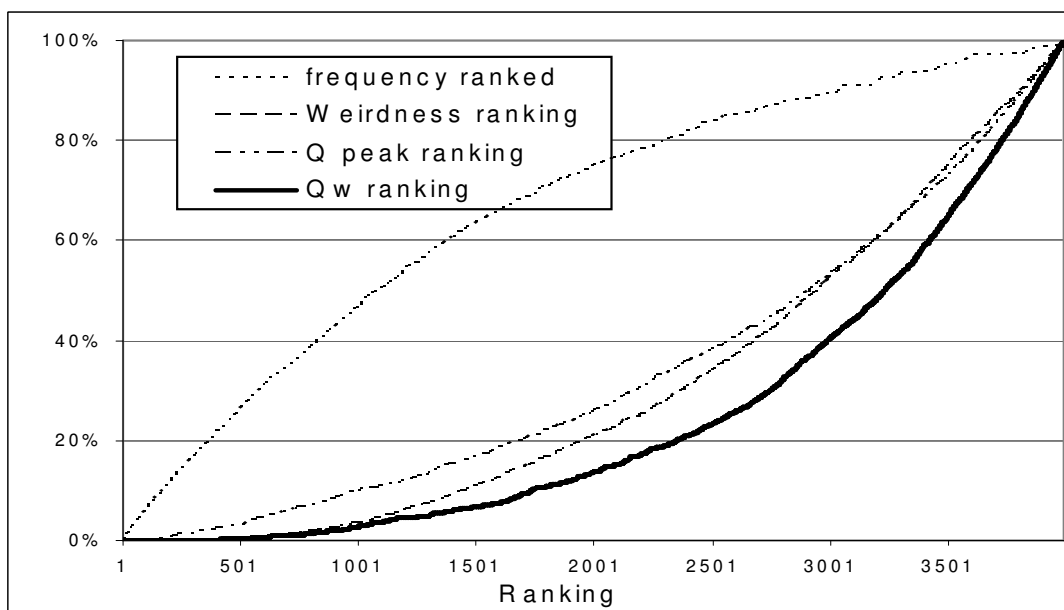


FIGURE 6: BNC top 2000 distribution comparisons

The BNC words have been pushed to the right by each ranking method, without ending up simply inverse frequency correlated. Using spearman’ coefficient, the correlation of Qw ranking and Frequency ranking is -0.30861, indicating little correlation, which is good.

Table 7 shows the top 10 and bottom 10 ranking words for each ranking method. BNC top 2000 words appear in bold.

Next we choose a cut off point for Qw and exclude all collocations with a Qw less than 1.2 (3606 words in total a 90% reduction). The frequency ranked table 8 shows the top 20 of the remaining 386 automatically selected collocations. Again the BNC top 2000 words appear in bold, these words would have been removed if these 2000 words had been used as a conventional stop list.

From this we can surmise Dasovich’s main interests/activities/correspondence in relation to “energy”. Similarly we can extract for Dasovich’s use of “Power”⁴ the top and bottom rankings (Table 9) and after auto stop word removal produce a top 20 ranking (Table 10) and similarly for “Contracts” (Table 11).

⁴ The nucleate “power” occurs 82,252 times, and the +/-5 token width vocabulary is 17,084.

Rank	Frequency	Weirdness	Q peak	Qw
1	the	dasovich	webmaster	inbox
2	to	jdasovic	castano	sempra
3	of	http	oom	dasovich
4	and	enron	fuelcell	jdasovic
5	in	ferc	rfb	pdt
6	a	calpine	blvd	haas
7	for	cpuc	dasovitch	iep
8	on	dynegy	eme	greenmountain
9	crisis	venturewire	federa	ferc
10	that	inbox	barretto	caem
~	~	~	~	~
3983	corner	towards	more	when
3984	filled	stood	when	got
3985	branch	man	southern	made
3986	popular	looked	fuel	er
3987	broken	er	calpine	she
3988	sitting	her	davis	did
3989	dismissal	britain	this	his
3990	accompanied	shook	made	him
3991	movement	men	utilities	shook
3992	stood	aye	as	her

TABLE 7: Dasovich "Energy" Top and bottom ranking comparisons

Rank	Word	Freq	Rank	Word	Freq
1	crisis	5053	11	dow	1446
2	california	4374	12	trading	1254
3	commission	4177	13	producers	1198
4	federal	3595	14	renewable	1141
5	enron	3399	15	costs	1118
6	regulatory	2996	16	markets	1095
7	companies	2368	17	jones	1064
8	duke	2042	18	sempra	1014
9	prices	1825	19	wholesale	1004
10	reliant	1638	20	senate	778

TABLE 8: Top 20 Dasovich "Energy" collocations after Auto Stop word removal.

Rank	Frequency	Weirdness	Q peak	Qw
1	the	dasovich	conf	charset
2	to	charset	plain	inbox
3	of	http	abag	hartsoe
4	and	enron	privileged	steffes
5	in	ferc	capitalists	etringer
6	for	calpine	hartsoe	newenergy
7	a	cpuc	thefts	dabhol
8	that	dynegy	orion	abag
9	on	steffes	amassing	dasovich
10	the	inbox	nth	cpuc
~	~	~	~	~
4278	bray	influences	a	sometimes
4279	assn	stolen	the	when
4280	dtrs	towards	as	got
4281	skeptical	hundred	that	programme
4282	sw	looked	over	his
4283	summ	er	on	then
4284	eix	her	between	er
4285	multiyear	church	back	her
4286	transwestern	eyes	because	war
4287	unsubscribe	programme	when	she

TABLE 9: Dasovich "Power" Top and bottom ranking comparisons

Rank	Word	Freq	Rank	Word	Freq
1	plants	6132	11	producers	1019
2	plant	3918	12	dwr	991
3	generators	2549	13	angeles	722
4	dabhol	2498	14	megawatts	712
5	wholesale	2464	15	dpc	705
6	enron	2209	16	megawatt	665
7	grid	1772	17	ferc	662
8	electric	1591	18	blackouts	569
9	purchases	1384	19	marketers	485
10	co	110	20	sellers	471

TABLE 10: Top 20 Dasovich “Power” collocations after Auto Stop word removal.

Rank	word	Freq	Rank	word	Freq
1	expiring	1133	11	stocks	1049
2	schedules	1130	12	sues	1042
3	nd	1125	13	analyze	1041
4	beforce	1123	14	billions	1039
5	retroactive	1122	15	finalize	1035
6	multiyear	1121	16	steven	1034
7	sacramento	1120	17	gen	1031
8	angelides	1119	18	byline	1030
9	sempra	1118	19	favorable	1029
10	outs	1052	20	fw	1028

TABLE 11: Top 20 Dasovich “Contracts” collocations after Auto Stop word removal

Measuring similarity of “energy” and “power” we might expect some degree of similarity of context, knowing a priori that Enron is an Energy/power company. Using formula 4 for “energy” and “power” for all tokens that match⁵, we get a similarity measure of 0.87, indicating a close similarity. As a comparison the similarity between “power” and “contracts”⁶ is 0.26, indicating a more distant relationship.

To complete the analysis we would need to determine whether the relationship is synonymous or hierarchical and here we may need to consider patterns of exclusions as identified above.

5.2 Much-more validation

To validate the approach we have performed the same operations using the Muchmore corpus⁷ on “case”⁸, “study”⁹ and “blood”¹⁰ and the results are presented in table 12. Here the Qw stop word cut off s set to 1.0. Again the BNC top 2000 words appear in bold, and we can again infer the dominant semantic associations.

Rank	patients word	freq	Study Word	freq	blood word	freq
1	underwent	154	Aim	320	pressure	237
2	chronic	140	present	239	flow	176
3	suffering	128	prospective	209	loss	75
4	postoperative	116	retrospective	116	arterial	60
5	undergoing	105	evaluate	107	cerebral	49
6	trauma	99	performed	56	vessels	42
7	tumor	95	randomized	50	gas	41
8	coronary	89	multicenter	33	samples	37
9	fractures	88	efficacy	22	peripheral	35
10	tumors	79	Shows	20	autologous	29
11	metastases	69	analyze	19	systolic	27
12	consecutive	68	Pilot	17	venous	23
13	artery	66	diagnostic	15	transfusions	23
14	atrial	65	analyzed	15	gases	17
15	prognosis	62	therapy	15	intraoperative	15
16	carcinoma	59	fractures	14	diastolic	14
17	resection	56	complications	13	uterine	13
18	malignant	52	longitudinal	12	glucose	13
19	critically	51	investigates	12	patch	12
20	according	51	demonstrates	12	cardiac	12

TABLE 12: MuchMore Top 20 “patients”, “study” & “blood” collocations: after auto stop word removal.

⁵ Stop words are not removed, their influence is naturally deprecated by small values for Q_peak and weirdness.

⁶ The nucleate “contracts” occurs 11,492 times, and the +/-5 token width vocabulary is 5,185.

⁷ The Springer Muchmore corpus of 7823 medical abstracts, vocabulary is 35033 types, total size is 1081124 tokens.

⁸ “patients” occurs 9,410 times, ranked 10 by frequency with a weirdness of 50.

⁹ “study” occurs 1,979 times, ranked 40 by frequency with a weirdness of 8.3

¹⁰ “blood” occurs 1,097 times, ranked 87 by frequency with a weirdness of 10

Using equation 4, for all matching words, similarity calculations indicate “patients” similarity to “study” (0.87) is high and “patients” similarity to “blood” (0.38) is low, however for “case” and “cases”¹¹ the similarity is 0.94 indicating more similarity as we would expect.

6 DISTRIBUTIONS OF DISTRIBUTIONS

We consider, finally, whether such results provide a set of predictable patterns, hence predictable keywords and collocations, and stop lists that can be used accurately with smaller proportions of the same corpus. In particular, we consider:

- Verticals: splitting the Enron corpus into sub-corpora, e.g. by user names - 150 separate users. Analysing and comparing use of language in these sub-corpora, with initial focus on “confidential” to demonstrate initial discriminance.
- Horizontals: marking out boundaries within texts (emails). A software component has been created to facilitate the identification and marking and collocation analysis is then used to provide the statistical evidence to reveal and investigate these less dominant contexts.
- Cumulative: For smaller collections we propose using the cumulative shape information of the collocates to infer entropy on the nucleate.
- Aggregate: To produce an aggregate stop list for general use, we propose to initially choose a nucleate at random and then to choose the next nucleate based on the collocation rankings and corpus collocation coverage. Once the net coverage overlap is greater than the unprocessed area, combine the nucleate collocation stop lists to build an aggregate stop list for the whole corpus.

These four orientations will assist both to identify other lower frequency distributions and to enable the accurate classification of body text versus banner.

Figure 7, below, demonstrates our rationale for verticals: we split collocation analysis across 4 quarters and investigated dominant patterns for the collocate “privileged” with “confidential”. One quarter appeared to produce unusual behaviour in contrast to the relatively similar distributions for the remaining quarters, evident in the difference in heights of the third bar at collocate positions L2 and R2. Either we were detecting different types of banners, or there was some other significant explanation such as legal discussions – for this.

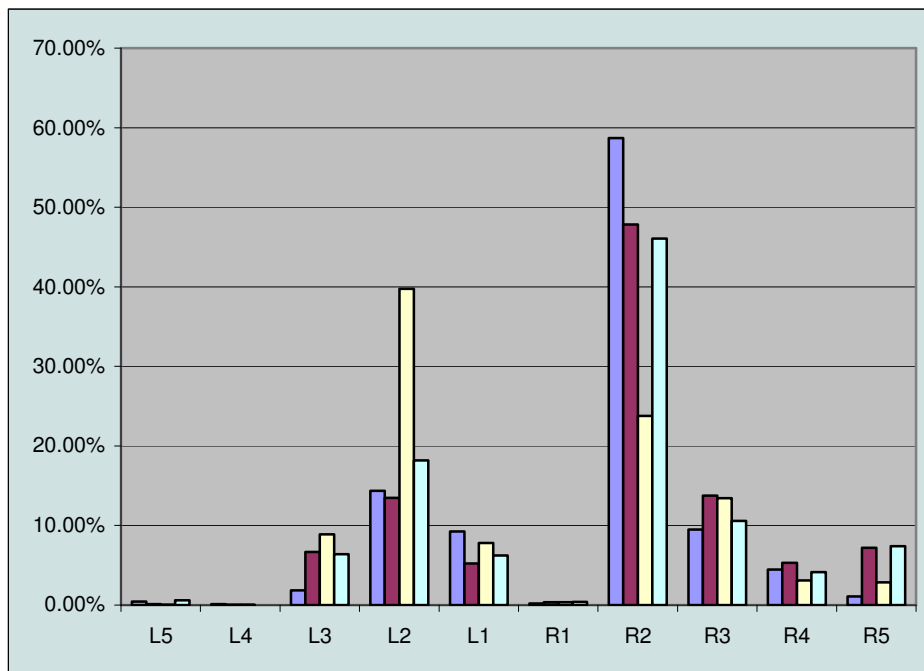


FIGURE 7: Vertical Spitting

7 CONCLUSIONS

Our aim is to be able to take an arbitrary text collection and automatically apply necessary and appropriate data cleansing techniques to it, have it produce its own stop word list, and provide a more robust ground for corpus analysis. This is necessary not only for email corpora such as that of Enron, but also for analysis of Blogs, Wikis, Web- Corpora, and other types of social or specialist collections.

Our approach has been evaluated and validated using publically available corpora, and tested against a BNC frequency list based stop list.

We have demonstrated that we can identify to some extent, certain kinds of information pollution within such corpora, and use statistical approaches to distributions for identifying stop words and lexical semantics. The approach attempts to remove assumptions relating to the text under analysis; it uses the collocation patterns associated with stop words combined with the relative frequency properties of stop words to derive the stop word list. We have demonstrated the ability to measure distributional lexical similarity and have compared “power” with

¹¹ not shown.

“energy” and “contracts”, showing that “power” is closely related to “energy” and only distantly related to “contracts”. However further work is required to bench mark the metric and understand how similarity is scaled in the range 0 to 1, and also in due course to extend and benchmark the semantic representation approach in this paper using the TOEFL® Test¹² in a similar manner to Bullinaria & Levy (2006).

BIBLIOGRAPHY

- [1] Bekkerman R; A. McCallum. & G. Huang, 2004, “Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora”. (Massachusetts).
- [2] Blei D M; A. Y. Ng; M. I. Jordan, 2003, Latent Dirichlet Allocation, *Journal of Machine Learning Research*.
- [3] Budanitsky A.; G. Hirst, 2006, *Evaluating WordNet-based Measures of Lexical Semantic Relatedness*, Association for Computational Linguistics, MIT Press.
- [4] Bullinaria J.A; J.P. Levy, 2006, Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study, *Behavior Research Methods*.
- [5] Cooke N; L. Gillam; A. Kondoz, 2007a, Best Kept Secrets of Corpus Linguistics, L2007 Birmingham.
- [6] Cooke N.; L. Gillam; A. Kondoz, 2007b, IP protection: Detecting Email based breaches of confidence, IAS2007 Manchester .
- [7] Fumera G; I. Pillai; F. Roli, 2006, Spam Filtering Based On The Analysis Of Text Information Embedded Into Images, *Machine Learning Research* 6: 2699-2720,
- [8] Gillam L.; K. Ahmad, 2005, Pattern mining across domain-specific text collections. *LNAI 3587*, pp 570-579.
- [9] Gillam L.; M. Tariq; K Ahmad, 2005, Terminology and the Construction of Ontology, *Terminology* 11(1), pp55-81. John Benjamins Publishing Company, ISSN 0929-9971; E-ISSN 1569-9994.
- [10] Hiemstra D.; F. de Jong, 2001, Statistical Language Models and Information Retrieval: natural language processing really meets retrieval, *Glott International* 5(8), pages 288-294, <http://www.glottinternational.com>.
- [11] Jabbari S.; B. Allison; D. Guthrie; L. Guthrie, 2006, Towards the Orwellian Nightmare Separation of Business and Personal Emails, *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 407–41.
- [12] Klimt B.; Y. Yang, 2004, The Enron Corpus: A New Dataset for Email Classification Research, *ECML 2004: 217-226 Language Technologies Institute, Carnegie Mellon University*.
- [13] Levy J.P.; J.A. Bullinaria, 2001, Learning Lexical Properties from Word Usage Patterns: Which Context Words Should be Used?
- [14] Li, W. (1992) “Random texts exhibit Zipf's-law-like word frequency distribution”. *IEEE Transactions on Information Theory* 38(6) pp1842-1845.
- [15] Lo. R.T; B. He; I. Ounis, 2005, Automatically building a stopword list for an information retrieval system, *The Journal on Digital Information Management: special issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR 2005)*
- [16] Manning, C. and Schütze, H. (1999) “Foundations of Statistical Natural Language Processing”. MIT Press. Cambridge, MA.
- [17] Pustejovsky, J.; P. Anick; and S. Bergler, (1993), “Lexical Semantic Techniques for Corpus Analysis”, *Computational Linguistics* 19(2): 331-358
- [18] Sinka M. P; D.W. Corne, 2003, Evolving Better Stoplists for Document Clustering and Web Intelligence Design and application of hybrid intelligent systems, 1015–1023.
- [19] Smadja F, 1993, Retrieving collocations from text: Xtract, *Computational Linguistics*, 19(1) pp143-178. Oxford University Press.
- [20] Trumbach C.C.; D. Payne, 2007, Identifying synonymous concepts in preparation for technology mining, *Journal of Information Science* 2007; 33; 660 originally published online Jun 14.
- [21] Watts R.J.; A.L. Porter; D. Zhu, 2002, Factor Analysis Optimization: Applied on Natural Language Knowledge Discovery, *Committee on Data for Science and Technology 2002: Frontiers of Scientific and Technical Data: Proceedings of the 18th International Conference CODATA 2002*
- [22] Wilbur W. J.; K. Sirotkin, 1992, The automatic identification of stop words, *Journal of Information Science* 18; 45
- [23] Yavuz, D. (1974) “Zipf's law and entropy” (Corresp.) *IEEE Transactions on Information Theory*, 20 (5). p650
- [24] Zipf, G.K. (1949) “Human Behavior and the Principle of Least Effort”. Hafner, New York.

AUTHOR BIOGRAPHIES

Neil Cooke BSc Computer & Control Systems Lanchester Polytechnic Coventry UK 1981 (now University of Coventry), Chartered Engineer 1987, Fellow of the Institute of Engineering Technology 2005. Studying part-time for a PhD, at the University of Surrey. The author has 18 years of experience in the field of information assurance engineering for the UK government’s National Technical Authority for Information Assurance. Prior to this the author was an avionics and marine systems engineer.

Lee Gillam PhD in Artificial Intelligence (Surrey, 2004); BSc in Mathematics and Computer Science (Surrey, 1995). Member of the British Computer Society (MBCS) and Chartered IT Professional (CITP). Currently a Lecturer in the Department of Computing at the University of Surrey. Previous publications and research in the areas of Ontology Learning, Metadata and Grid Computing Systems.

¹² <http://www.ets.org/portal/site/ets/>