

Extracting Significant Words from Corpora for Ontology Extraction

Dileep Damle
Knowledge Media Institute, The Open University
Milton Keynes, MK7 6AA, UK
d.g.damle@open.ac.uk

Abstract

This paper reports a technique for Knowledge Extraction using Natural Language Processing for the purposes of semi-automatic Ontology learning. Determination of significant words in a relevant collection of text is an important first step in building ontologies from natural language text. However, terminology identification is also a slow and expensive process requiring terminological and domain expertise. We report experiments with three different document collections comparing word frequency distributions over documents against a reference corpus representing more general subject matter.

1. Introduction

This work is part of a larger project to build ontologies, or knowledge models, by processing a collection of documents (the domain corpus) relevant to some area of knowledge - a knowledge domain. Such ontologies have potential uses in a number of applications such as concept based web navigation. Currently, Ontology construction is a slow and expensive manual process requiring expertise in the domain and in ontological engineering. Both kinds of experts may be scarce and this provides the motivation for this research.

Automating the ontology building process requires sources of knowledge about the domain to be modelled. In general, existing glossaries, terminologies, data models, ontologies of parts of the domain as well as text sources of various kinds all may be useful resources for ontology construction (See Maedche and Staab[1]). However for many domains, the only sources that exist are textual. For such domains, ontology learning from natural language text may be the only available option.

If an ontology is to be semi-automatically constructed from a domain corpus, the corpus itself or the constructors of the corpus, implicitly define the knowledge domain. Domain experts and users may refine such a definition later by editing the ontology.

The process of constructing an ontology from a corpus may be viewed as consisting of three distinct steps. The first step is to identify some adequate number of the most important words, or terms in the domain corpus. The terms may be seen as proxy for the concepts in the domain. Although, the interest here is in the important domain concepts rather than terms, initial identification of terms is still important. Much of the progress in this has been made by lexicography and taxonomy oriented researchers who are primarily interested in nouns that represent entities. For ontology building we intend also to exploit other categories of content words and their co-occurrences.

In the second step, hypotheses about the properties and relationships between these terms may be formulated from analysis of fragments of text that contain these terms. The final step is evaluation of these hypotheses and construction of

an ontology from the hypotheses that are accepted.

This paper reports a part of the work in the first phase, term identification. We find that comparing the frequencies of lemmas, within their coarse part of speech categories (Nouns, Verbs, Adjectives and Adverbs) with those within a reference corpus provides promising results.

2. Previous Research

The simplest approach to identifying terms is to rank words according to their frequencies in the text, disregarding function words which occur with high frequencies. More sophisticated approaches compare relative frequencies of words within the corpus of interest with those in a reference corpus representing the general language. For example, Drouin [2] constructed a reference corpus from other corpora in related but different domains. If an existing corpus representing the host language is readily available, then this may be used. The statistical confidence level in the frequency differences is often evaluated by assuming that the Normal distribution can be used.

If the relative frequencies of the important words are high enough to permit the assumption of the Normal distribution for hypothesis testing, then well and good. However, this raises some problems.

First, Church and Gale [3] and Church [4] and have shown that frequency distributions of topical words are more akin to Poisson Mixtures than to Normal Distributions.

Second, some important domain-specific words may also occur with very low probabilities, but these will be mixed among low frequency words from the general language too. How to tell the difference? We show that using a test based on Poisson distribution provides very good results. It has not

been possible to perform a direct comparison between the methods due to the lack of a standard benchmark test.

3. The Method

We wanted to find out which kinds of text sources provide good results and collected 3 corpora in scientific/technical domains - a collection from the word-wide web, a collection of papers from a single journal and an e-book. These were our 'Target Corpora' or TCs. The details are in Table -1. The British National Corpus - World Edition (BNC) was used as a reference corpus (RC). We identified the documents belonging to the genre "Informative: Natural & pure science" (wridom2) and these constituted our reference corpus. We extracted the same information for the BNC corpus as for our experimental corpora. We did not apply the Part-of-speech tagger again, but translated the POS tags in the BNC into Penn tags at a later stage as we were only interested in the very coarse tags of 'Noun', 'Verb', 'Adjective', 'Adverb' and 'Other'.

After typical pre-processing including part-of-speech tagging and lemmatisation, we counted the word frequencies for each separate document. The corpus named CP contained a variety of documents from a variety of sources and required much effort to overcome variation in orthographic styles.

The next step was to apply the collocation extraction techniques of Frantzi and Ananiadou [5]. This step only involved the three TCs, as the BNC already identifies many collocations. The collocations thus identified were substituted back into the documents.

It is worth pointing out that, for each word we encountered, we created a lemma, and distinguished it by its original POS tag. Thus, 'cut' and 'cutting' as nouns were identified as

‘cut/N’ while the same words encountered in their verb forms were identified as ‘cut/V’. In what follows we refer to these POS differentiated lemmas as lemmas. The aggregation of the data into lemmas and coarse parts of speech categories was intended to boost

the relative word occurrence frequencies. Relative within-document frequencies were also further boosted by defining with respect to the total number content words only rather than all words including function words.

Table 1 – Corpora Details

Corpus	Documents	Sentences	Words	Content words	Comments
CP	216	32463	513091	294471	This was gathered by running a Google query ‘Climate Prediction’ 3 times selecting pdf, MSWord and HTML documents respectively and then downloading from the resulting links. Some broken links resulted in less than 300 documents being in the corpus.
JC	48	17072	424817	230684	This consists of 48 articles from Journal of Climate (1998-2005)
Neuro	8	6258	119816	64713	An ebook – ‘Neuroscience of psychoactive substance use and dependence’ (WHO, Geneva). This was obtained as a single document, but split into chapters. The contents page and the index was omitted from the processing. The index allows comparison with the experimental results.
BNC					British National Corpus Word Edition. Only documents with classification wridom=2 meaning “Informative: Natural & Pure Science” were used.

Our data can be defined as follows:-

For each Lemma l , and corpus C we have collection of vectors:

$V_i = [f_i, l_i]$; where f_i = no of occurrences of l in every document i in the corpus C in which that lemma occurs. l_i is the number of lemmas in that document (thus not counting non-content words). This way of defining the relative word frequency raises the value by ignoring the content words.

The distributions of these relative frequencies implied by the pairs in the vectors V_i were compared between the TCs and the RC. The next section provides the mathematical details.

4. The Mathematical Details

Given two corpora RC and TC, for any given lemma occurring in both corpora:

In document i of RC, Poisson parameter for the lemma is λ_i and the document length is l_i . Thus frequency of lemma within the document is $f_i = \lambda_i \cdot l_i$.

x_i is the number of times the lemma occurs in i th document.

$$f(x_i) = \frac{e^{-l_i \lambda_i} (l_i \lambda_i)^{x_i}}{x_i!}$$

For the complete RC corpus:

$$\text{Likelihood: } \varphi_{RC} = \prod_{i=1}^N f(x_i) = \prod_{i=1}^N \frac{e^{-l_i \lambda_i} (l_i \lambda_i)^{x_i}}{x_i!}$$

For the complete TC corpus:

$$\text{Likelihood: } \varphi_{TC} = \prod_{j=1}^M f(y_j) = \prod_{j=1}^M \frac{e^{-l_j \lambda_2} (l_j \lambda_2)^{y_j}}{y_j!}$$

H0: There is no real difference between the two corpora for this lemma, we have a single λ .

Likelihood for the combined corpus:

$$\varphi_{TOT} = \varphi_{RC} \times \varphi_{TC}$$

$$\log(\varphi_{RC}) = -\sum l_i \lambda_1 + \sum x_i \ln(l_i \lambda_1) - \log(x_i!)$$

$$\frac{\partial \log \varphi_{RC}}{\partial \lambda_1} = -\sum l_i + \sum \frac{x_i}{\lambda_1}$$

$$\lambda_1 = \frac{\sum x_i}{\sum l_i}$$

Setting this to 0 :

Hence the log maximum likelihoods are:

$$\log(\varphi_{RC}) = -\sum l_i \frac{\sum x_i}{\sum l_i} + \sum x_i \ln\left(l_i \frac{\sum x_i}{\sum l_i}\right) - \sum \log(x_i!) \quad (1)$$

$$\log(\varphi_{TC}) = -\sum l_i \frac{\sum y_i}{\sum l_i} + \sum y_i \ln\left(l_i \frac{\sum y_i}{\sum l_i}\right) - \sum \log(y_i!) \quad (2)$$

$$\log(\varphi_{TOT}) = -\sum l_i \frac{\sum z_i}{\sum l_i} + \sum z_i \ln\left(l_i \frac{\sum z_i}{\sum l_i}\right) - \sum \log(z_i!) \quad (3)$$

Log likelihood ratio: (1)+(2)-(3) =

$$\left(\sum x_i \ln\left(l_i \frac{\sum x_i}{\sum l_i}\right) + \sum y_i \ln\left(l_i \frac{\sum y_i}{\sum l_i}\right) - \sum z_i \ln\left(l_i \frac{\sum z_i}{\sum l_i}\right) \right)$$

Compare **twice this ratio** with a χ^2 distribution on 1 degree of freedom and reject if the likelihood ratio is too large.

4. Results and discussion

The results are presented in Tables 2a-d and Table 3. For the sake of brevity, Tables 2a-d present only the top 20 extractions in each part of speech category respectively. The full extraction lists contain many terms that intuitively appear correct. However, Table 3, which shows the extraction rates as percentages of the words actually compared between the TC and RC for each of the three corpora; better represents the degree of success achieved. This is despite the number of issues we identify in this discussion.

Tables 2a-d also demonstrate that the Poisson approach works with quite low relative frequencies, where the Normal distribution approximation would not be valid due to low absolute frequencies.

It is noticeable that the percentage extraction rates fall off as you move from CP, through JC to the Neuro corpus. This is counter-intuitive as we expected to get better results from the more definitive documents than from a simple collection based on a Google

Table 2a – Top 20 Nouns

	CP				JC				Neuro			
	lemma	CHI_sq	ABS	Rel	lemma	CHI_sq	ABS	Rel	lemma	CHI_sq	ABS	Rel
1	climate	5176.3	523	0.002834	model	2181.9	426	0.002743	dependence	1914.7	146	0.003403
2	model	1957.7	388	0.002102	<i>fig</i>	1734.9	1080	0.006954	substance	1610.1	105	0.002447
3	prediction	1834.6	282	0.001528	anomaly	1663.3	163	0.001050	<i>al</i>	1053	1028	0.023958
4	forecast	914.9	494	0.002677	wind	1342.9	95	0.000612	use	815.4	118	0.002750
5	weather	537.2	141	0.000764	precipitation	1161.6	201	0.001294	drug	565.3	160	0.003729
6	precipitation	466.8	196	0.001062	cloud	943.7	142	0.000914	alcohol	546.1	139	0.003239
7	drought	453	129	0.000699	response	870.5	195	0.001256	receptor	397.6	66	0.001538
8	<i>information</i>	420.9	190	0.001029	flux	795.2	81	0.000522	effect	275.5	77	0.001795
9	ocean	380.4	194	0.001051	mean	668.2	326	0.002099	withdrawal	273.8	49	0.001142
10	monsoon	348.9	66	0.000358	variability	629.7	85	0.000547	disorder	219.8	43	0.001002
11	center	340	48	0.000260	<i>al</i>	605.5	1167	0.007514	smoking	197.3	70	0.001631
12	<i>user</i>	335.2	165	0.000894	simulation	572.1	112	0.000721	depression	191.6	72	0.001678
13	variability	331.6	102	0.000553	feedback	563.8	81	0.000522	brain	183	57	0.001328
14	decision	320.1	142	0.000769	event	529.6	63	0.000406	research	168.8	50	0.001165
15	impact	319.6	123	0.000666	storm	501.4	50	0.000322	dopamine	165	60	0.001398
16	outlook	319.4	44	0.000238	climate	499.5	50	0.000322	treatment	148.2	91	0.002121
17	circulation	272.4	52	0.000282	ocean	470.3	169	0.001088	schizophrenia	146.3	83	0.001934
18	season	244.8	147	0.000796	trend	445.4	100	0.000644	nicotine	127.3	87	0.002028
19	rainfall	241.4	126	0.000683	heat	401.6	57	0.000367	polymorphism	121.1	6	0.000140
20	moisture	230.9	34	0.000184	rainfall	391.2	69	0.000444	consent	107.5	43	0.001002

Table 2b - Top 20 Verbs

	CP				JC				Neuro			
	lemma	CHI_sq	ABS	Rel	lemma	CHI_sq	ABS	Rel	lemma	CHI_sq	ABS	Rel
1	predict	380	335	0.001815	force	1672.1	884	0.005692	induce	140	109	0.002540
2	improve	370.3	347	0.001880	couple	642.5	380	0.002447	increase	126.2	156	0.003636
3	include	368.7	576	0.003121	observe	530.3	599	0.003857	associate	102.9	126	0.002937
4	<i>will</i>	323	1404	0.007607	<i>have</i>	391.3	1236	0.007959	involve	83.2	122	0.002843
5	base	303.9	424	0.002297	mix	372.1	234	0.001507	relate	83.1	107	0.002494
6	provide	296.4	599	0.003246	simulate	354.7	256	0.001648	alter	78.5	59	0.001375
7	couple	274.3	194	0.001051	average	307.8	352	0.002267	inform	76.4	34	0.000792
8	<i>shall</i>	263	301	0.001631	<i>show</i>	264.9	1082	0.006967	<i>shall</i>	65.8	77	0.001795
9	develop	261.2	517	0.002801	indicate	226.1	462	0.002975	reinforce	57.7	42	0.000979
10	be	253.3	9822	0.053219	<i>mean</i>	226.1	326	0.002099	learn	46.5	55	0.001282
11	repeat	175.1	100	0.000542	associate	218.2	362	0.002331	repeat	42.4	40	0.000932
12	force	168.3	149	0.000807	curl	210	80	0.000515	develop	38.5	102	0.002377
13	relate	165.7	273	0.001479	base	209	367	0.002363	mediate	34.8	40	0.000932
14	need	146.7	437	0.002368	<i>decrease</i>	177.9	241	0.001552	prolonged	34.7	22	0.000513
15	issue	146.4	165	0.000894	induce	177.3	191	0.001230	<i>decrease</i>	33.6	48	0.001119
16	compute	142.3	93	0.000504	enhance	165.9	156	0.001004	treat	31.1	53	0.001235
17	simulate	125.5	87	0.000471	compute	156.7	121	0.000779	discuss	30.6	58	0.001352
18	<i>using</i>	125.1	397	0.002151	estimate	153.5	298	0.001919	lead	30.2	84	0.001958
19	observe	123.2	248	0.001344	<i>increase</i>	134	420	0.002704	include	29.9	92	0.002144
20	<i>have</i>	122.9	1809	0.009802	fix	133.9	120	0.000773	condition	29.5	22	0.000513

Table 2c – Top 30 Adjectives

CP				JC				Neuro				
	lemma	CHI_sq	ABS	Rel	lemma	CHI_sq	ABS	Rel	lemma	CHI_sq	ABS	Rel
1	seasonal	948.3	243	0.001317	atmospheric	775.9	80	0.000515	genetic	243.1	21	0.000489
2	regional	717.5	113	0.000612	tropical	695.6	57	0.000367	ethical	122.8	29	0.000676
3	normal	637.5	216	0.001170	due	546.6	393	0.002531	psychoactive	87	6	0.000140
4	global	593.4	100	0.000542	annual	494.8	61	0.000393	dependent	84.1	39	0.000909
5	tropical	343	34	0.000184	equatorial	379.7	23	0.000148	chronic	77.8	7	0.000163
6	operational	325.1	45	0.000244	maximum	377.9	119	0.000766	pharmacological	76.5	4	0.000093
7	atmospheric	288.3	63	0.000341	global	367.7	70	0.000451	<i>due</i>	65.8	43	0.001002
8	due	221.6	149	0.000807	cold	362.7	88	0.000567	cognitive	56.4	6	0.000140
9	monthly	221.1	83	0.000450	vertical	347.8	28	0.000180	mental	45.5	9	0.000210
10	statistical	211.6	36	0.000195	warm	329.1	89	0.000573	potential	41.4	33	0.000769
11	agricultural	199.5	38	0.000206	seasonal	325.7	23	0.000148	psychiatric	40.8	10	0.000233
12	warm	183.2	59	0.000320	eastern	311.6	22	0.000142	therapeutic	40.3	7	0.000163
13	meteorological	174.3	60	0.000325	extreme	301.5	29	0.000187	specific	38.2	12	0.000280
14	daily	161.8	44	0.000238	western	270.4	21	0.000135	human	36.2	27	0.000629
15	initial	155.4	20	0.000108	anomalous	261.5	25	0.000161	rewarding	35.7	4	0.000093
16	local	152	60	0.000325	negative	258.4	80	0.000515	twin	34.4	15	0.000350
17	<i>available</i>	147.3	230	0.001246	daily	253.1	12	0.000077	negative	33.6	5	0.000117
18	<i>current</i>	143.3	55	0.000298	significant	250.2	147	0.000947	solvent	31	5	0.000117
19	potential	140	63	0.000341	positive	245.1	94	0.000605	legal	30.9	7	0.000163
20	probabilistic	139	21	0.000114	solar	238.9	58	0.000373	individual	30.7	58	0.001352

Table 2d – Top 20 Adverbs

CP				JC				Neuro				
	lemma	CHI_sq	ABS	Rel	lemma	CHI_sq	ABS	Rel	lemma	CHI_sq	ABS	Rel
1	<i>well</i>	270.7	330	0.001788	statistically	189.7	102	0.000657	<i>also</i>	168.1	267	0.006223
2	<i>also</i>	127.7	566	0.003067	respectively	184	239	0.001539	<i>well</i>	34	63	0.001468
3	<i>longer</i>	100	67	0.000363	<i>also</i>	143.5	730	0.004700	ethically	28.8	13	0.000303
4	<i>especially</i>	85.2	112	0.000607	eastward	143.3	91	0.000586	legally	28.7	13	0.000303
5	<i>generally</i>	64.4	84	0.000455	northward	127	88	0.000567	widely	25.6	26	0.000606
6	<i>ahead</i>	63.5	37	0.000200	<i>well</i>	121.4	269	0.001732	directly	21.2	26	0.000606
7	<i>directly</i>	60.4	54	0.000293	significantly	103.4	149	0.000959	<i>longer</i>	17.3	13	0.000303
8	<i>currently</i>	59.9	90	0.000488	<i>somewhat</i>	87.8	81	0.000522	<i>most</i>	16.7	81	0.001888
9	<i>most</i>	56	269	0.001458	<i>generally</i>	79	135	0.000869	<i>freely</i>	15.3	10	0.000233
10	<i>effectively</i>	50.1	41	0.000222	primarily	64.3	58	0.000373	potentially	13.6	15	0.000350
11	<i>forward</i>	48.1	35	0.000190	approximately	61	74	0.000476	critically	13.1	10	0.000233
12	<i>rather</i>	45.8	74	0.000401	relatively	60.4	114	0.000734	biologically	12.6	8	0.000186
13	<i>somewhat</i>	43.3	37	0.000200	seasonally	55.9	36	0.000232	<i>rapidly</i>	11.9	14	0.000326
14	<i>typically</i>	40	36	0.000195	<i>here</i>	55.6	231	0.001487	<i>currently</i>	11.5	23	0.000536
15	<i>very</i>	38.3	223	0.001208	<i>longer</i>	53	53	0.000341	genetically	11	10	0.000233
16	<i>already</i>	37.9	62	0.000336	downwind	51.1	19	0.000122	<i>previously</i>	9.9	15	0.000350
17	<i>underway</i>	36.8	29	0.000157	<i>especially</i>	49.8	112	0.000721	<i>alone</i>	9.1	12	0.000280
18	eastward	36.5	26	0.000141	<i>rather</i>	48.6	111	0.000715	<i>ultimately</i>	8.9	8	0.000186
19	even	36	163	0.000883	<i>slightly</i>	47.7	103	0.000663	<i>naturally</i>	8.5	8	0.000186
20	westward	35.9	24	0.000130	<i>prior</i>	44.8	59	0.000380	<i>mostly</i>	8.3	9	0.000210

Table 3 – Percentage of tested Terms significant at P = 0.01

	CP			JC			Neuro		
	Number compared	significant at P=0.01		Number compared	significant at P=0.01		Number compared	significant at P=0.01	
		number	%		number	%		number	%
Nouns	755	572	76%	1433	723	50%	959	388	40%
Verbs	502	410	82%	823	424	52%	549	152	28%
Adjectives	181	166	92%	495	311	63%	253	109	43%
Adverbs	147	134	91%	334	193	58%	190	56	29%

Notes for Tables 2a-2d:

1. CHI_sq column has the hypothesis testing statistic. The ABS column shows the total number of occurrences in the whole of the TC and Rel columns shows frequency of occurrence of the lemma over the whole corpus relative to all lemmas in that part of speech.
2. For significance testing, **twice** the value in the CHI_sq column should be compared with the critical values in the CHI Squared Tables. The critical values for 1 degree of freedom are 3.841 for P = 0.05 and 6.635 for P = 0.01
3. We have **subjectively** marked these tables to indicate where we agree with the results – lemmas in **bold** type are the ones we think are domain specific, and the others are in *italics*.

Notes for Table 3

1. The numbers in the ‘Number compared’ column reflects the fact that not all terms in a TC were also in the RC. Such terms could not therefore be compared and are excluded here.

Table 4 -Words extracted with (P < 0.01) compared with the words in the index

Word	N	V	Adj	Adv
abuse	53			
acetylcholine	222			
action	48			
acute			25	
agonist	214			
alcohol	6			
amphetamine	78			
analysis	151			
animal	208			
antagonist	173			
antisocial			44	
area	231			
cellular			80	
chemical	375			
chronic			5	
clinical			62	
cocaine	26			
coercion	135			
conditioning	85			
confidentiality	197			
consent	20	47		
consumption	60			
control		42		
cortex	29			
cultural			101	
deficit	99			
dependence	1			

depression	12			
disorder	10			
dopamine	15	84		
drug	5			
element	210			
environmental			73	
epidemic	299			
epidemiological			28	
ethical			2	
exposure	183			
frontal			26	
function		132		
gene	181			
genetic			1	
genetics	51			
genotype	105			
global			51	
glutamate	234			
heroin	88			
high			57	
human	127		14	
hyperactivity	330			
hypothesis	56			
illness	253			
imaging	259			
immunotherapy	93			
in	174			
incentive	58			

independent			86	
individual	216		20	
induction	336			
instrumental			38	
intervention	271			
justice	213			
legal			19	
legally				4
linkage	76			
mechanism	68			
metabolism	185			
motivation	94			
negative			17	
neurobiology	270			
neuronal			23	
opiate	209			
opioid	34			
organization	165			
other	131		29	
pathway	360			
peptide	353			
personality	147			
pharmacological			6	
plasma	237			
positive			67	
potential			10	
Rest omitted for lack of space				
" 126 Significant words out of 597 unique words in index "				

search. Two explanations suggest themselves.

First possibility is that the relative size of the domain specific vocabulary in each corpus may be different. In particular, CP was constructed blindly based on a simple Google search. The documents came from a variety of sources and authors and with little or diverse editorial control. This may have led to a much more diverse domain vocabulary. By contrast The JC corpus is built from a single journal with the controls and consistency of language use that is implied. The Neuro corpus is constructed from separating the chapters of a single book which has many authors, but probably a single purpose and a strong editorial control. So, a possible explanation may be that increasing the degree of control lead to smaller specialist vocabularies.

The alternative explanation may be based on variation in corpus granularity at the document level. If terms or topical words occur in bursts as noted by Church and Gale, then larger documents will be more likely include Poisson mixtures than many smaller documents, in which case our model which assumes a single Poisson frequency for each word within a document does not fit as well. It is also possible that both explanations are correct.

Another observation that might be made is that there is no obvious difference between the extraction rates for different parts of speech. This suggests that ontology building may find it useful to exploit all content words rather than just nouns and noun phrases with verbs providing evidence of processes, and adjectives and adverbs providing evidence of properties. Nouns only provide evidence of things that are.

In addition to the extraction lists which result from the comparison between TC and RC frequencies, we also produced lists of noun phrase collocations using

the Frantzi and Ananiadou method. It might have been more appropriate to defer the collocation extraction process to a later stage because we found little occurrence of these domain relevant collocations occurring in the RC. This may be a problem because collocations composed of apparently normal language words such as *Climate* and *Prediction* do not occur in the RC and we have no automatic method for determining their topicality.

We made an initial attempt at evaluation of the results using the Neuro corpus which provides a large index. We took all the words in the index after some simple initial pre-processing to remove usages of ‘, see also’, ‘and’ etc. After removing all the duplicates from this list, we compared it with the words we had determined as significant with a P value of 0.01. These results are in Table 5. Only some words are shown and we found that 21% of the words in the index were in our list. This appears quite low and is due to the fact that many of the index entries are phrases, potentially corresponding to the collocations which had been excluded from the analysis for lack of evidence in the RC. The Comparing collocations with a C/NC value above 2.0, from our list with the index, we found that three two word collocations (‘substance dependence’, ‘psychoactive substance’ and ‘psychoactive substance’) took part in 16 other collocations preventing these matches from occurring. We also found that the phrase ‘without dependence’ that we detected as a collocation occurs as a second level item (modifying the top level index item under which it appears) in the index 26 times. When corrected for these two discoveries, we appear to have detected 28% of the index items. However, we do not take this evaluation seriously, because and index is not really a list of all the significant words, nor is it a purely objective artefact – its content depends

on both the author's own bias as to what is important and the needs of the readership as perceived by the author.

A comparison with a glossary or an index has certain limitations. The indices and glossaries are usually ordered alphabetically –no indication of the importance of index or glossary items is provided. In a few cases some limited information about order of importance is available when an indented index is provided (The Neuro index is such), but this is within a flat alphabetical structure and the structural depth may be no more than 3 at best. In the case of an index the number of pages referred to may well be another measure of importance, but that could also be for other reasons. Ultimately, we do not believe that there is a fully objective test for this kind of work, the importance of words is in the mind of the intended user and not necessarily a characteristic of the domain itself.

5. Future work

We are at a stage where the extraction lists are adequate for supporting the ontology learning experiment that is the main focus of our research. However, an objective evaluation technique for the work reported here would be useful. We plan to carry out with further corpora examples, particularly looking for any that provide good evaluation opportunities.

We would also like to see if the size and granularity of the corpus explains the variation in extraction rates by repeating the extraction over differently sized randomly selected subsets of the same corpus. If the explanation lies in size and granularity rather than in editorial control and multiple authorship, then there may be a case for experimenting with inter-occurrence gaps as the test statistic.

The inter-occurrence gap may be a better statistic because it does not involve document length. It also allows burstiness to be dealt with naturally. Without fixing a window size within which to calculate the frequency or the distribution of the gap sizes, the very large gaps can simply be ignored. Ignoring the gap sizes larger than some threshold value may act as a high-pass filter in respect of the frequencies.

It may also be better to use two RCs in sequence. The first RC would be representative of a much bigger part of the language (for example using the whole of the BNC). This would allow some comparison of many more words. The remaining words would then be genre and domain specific. A second specialising RC would be constructed by adding together corpora from adjacent domains. For example, for the Neuro corpus, this would be a mixture of Psychiatry, Neurochemistry, Pharmacology, etc. Such an adjacent RC would differentiate the specialist domain words from adjacent domain words and genre specific words.

However, our main focus will now be on extraction of entities and relationships. The next step in this is to extract sentences from the corpus based on the occurrence of the most significant lemmas. We aim to rank the sentences according to the CHI-squared values for single lemmas and C/NC values for collocations, also taking into account the occurrence of multiple target lemmas. We expect this to order the sentences by degree of informativeness about the principal terms. We will then explore techniques for making inferences about entities and relationships from these sentences, using available resources such as WordNet and FrameNet for exploiting what coded knowledge is available for the English language in its general form.

7. Acknowledgements

This work was carried out with the support from the EPSRC and the Open University. I would also like to acknowledge the advice of my supervisors Anne DeRoeck, John Domingue and Victoria Uren. I also gratefully acknowledge the formulae for the maximum likelihood calculation provided by Professor Paul Garthwaite.

6. References

1. Maedche, A. and S. Staab, *Ontology learning for the Semantic Web*. Intelligent Systems, IEEE [see also IEEE Expert], 2001. **16**(2): p. 72-79.
2. Drouin, P. *Detection of Domain Specific Terminology Using Corpora Comparison*. in *fourth international Conference on Language Resources and evaluation*. 2004. Lisbon, Portugal.
3. Church, K.W. and W.A. Gale, *Poisson Mixtures*. Journal of Natural Language Engineering, 1995. **1**(2): p. 163-190.
4. Church, K.W. *Empirical estimates of adaptation: the chance of two Noriegas is closer to $p/2$ than p^2* . in *17th International Conference On Computational Linguistics*. 2000. Saarbrucken, germany: Association for Computational Linguistics Morristown, NJ, USA.
5. Frantzi, K.T. and S. Ananiadou. *Extracting Nested Collocations*. in *16th International Conference on Computational Linguistics*. 1996. Copenhagen, Denmark: ACL, Morristown, NJ, USA.