COLLOCATION ANALYSIS

Jeffrey How - [redacted]

COMP6781

Table of Contents

Experiment Background 1
Input Files, Tools & Libraries 1
Output 1
Programming 2
Word 2
Bigram 2
POSValidator 2
BigramDatabase 2
Driver 2
Analysis & Results 3
Background 3
Odd Couples: Headings & Textbook conventions 3
T-Score Analysis 4
χ^2-Score Analysis 4
Areas of Improvement 5
More refined POS tag set 5
Updated text corpora 5
Larger sample size 6
More comprehensive data analysis 6
Experimentation 6
Program Execution Instructions 6
Appendix 7
Exhibit I: Biology – Frequency Top Ten 7
Exhibit 2: Biology – T-Score Top Ten 7
Exhibit 3: Biology – χ^2-Score Top Ten 7
Exhibit 4: Biology – Frequency Bottom Ten 8
Exhibit 5: Biology – T-Score Bottom Ten 8
Exhibit 6: Biology – χ^2-Score Bottom Ten 8
Exhibit 7: Politics – Frequency Top Ten 9
Exhibit 8: Politics – T-Score Top Ten 9
Exhibit 9: Politics – χ^2-Score Top Ten 9

Exhibit 10: Politics – Frequency Bottom Ten 10
Exhibit 11: Politics – T-Score Bottom Ten 10
Exhibit 12: Politics – χ^2-Score Bottom Ten 10
References 11

Experiment Background

The purpose of this document is to provide the reader with a description of the attached collocation identifier program. Additionally, an analysis of its experimentation results will be provided.

Input Files, Tools & Libraries

- Corpora from the following three domains can be found in the '*corpora*' directory.

Domain	# of Sources	Total # of Words
Politics	8	168,509
Philosophy	1	128,761
Biology	1	68,765

- *inputFiles\PennTagSet_Selection2.txt*: This file includes the **selection** of **Penn Treebank POS tags** deemed **acceptable** for **collocation tag patterns**.
- *inputFiles\ignore_list.txt*: Despite POS filtering, a few common content-lacking words were still included since they had acceptable tags. This file includes a handful of words to filter out since they tend to lack content. Some examples include: 'had', 'has', 'is', 'be', 'are', etc.
- *Stanford POS Tagger*[1]
- *Apache Commons Mathematics Library*[2]

Output

- *outputFiles\output_[**domain_name**].txt*: The output text file lists the top 100 bigrams in order of frequency. For each bigram, the **frequency**, **t-score** and **chi-score** are provided. The parameters for the scores were solely based on word and bigram frequencies. Further, analysis on these results was performed in Excel, where the lists are more easily sortable. See 'Results' section.

[1] http://nlp.stanford.edu/software/tagger.shtml
[2] http://commons.apache.org/proper/commons-math/

Programming

The Java program provided consists of the following classes:

Word

A word consists of two components: a case-sensitive lexeme and a part of speech. In the context of the regarded program, these elements distinguish a unique word (as opposed to the lexeme alone).

Bigram

A bigram is a class that holds two word objects (word1 and word2).

POSValidator

This singleton helper class is used to validate if a given word's part of speech is acceptable according to its input file (*inputFiles\PennTagSet_Selection2.txt*).

BigramDatabase

The Bigram Database uses the TreeMap[3] class to store the bigrams and words of an input file. It uses the POSValidator to filter out unacceptable bigrams. Additionally, it has the ability to output the frequency, **t-score** and **χ^2-score**[4] results of the stored bigrams.

Driver

This program lets the user choose a text file to read. Afterwards, it POS tags the file using the **Stanford POS Tagger**[5]. Once the file is tagged, the program uses a BigramDatabase object to identify potential collocations.

[3] http://docs.oracle.com/javase/6/docs/api/java/util/TreeMap.html
[4] http://commons.apache.org/proper/commons-math/apidocs/org/apache/commons/math3/stat/inference/ChiSquareTest.html
[5] http://nlp.stanford.edu/software/tagger.shtml

Analysis & Results

Background

Analysis was performed by looking at the results two particular domains: biology and politics. Biology was chosen due to its highly terminological context. Thus, expected collocations would be tend to be scientific or medical expressions. As seen in Exhibits 1-6, this expectation turned out to be true, were concepts like 'spinal cord' or 'nervous system' were identified.

Politics, on the other hand, has a much wider scope. It can relate to topics such as the economy, taxes, abortion, law, human rights, etc. Therefore, to experiment with a large scope of politics, multiple sources were taken from multiple political articles or texts. Please refer to the References section for text sources.

Frequency, T-Score and χ^2-Score were calculated for the 100 most-frequent bigrams. The text files with the results were opened in Excel for analysis (delimited by tabs). The T-Score ranks and χ^2-Score ranks were then compared to the Frequency ranks to identify large variations. Note that the T-Score and χ^2-Score rankings are approximate rankings. They were based by rearranging the top 100 frequent bigrams, not the entire collection of all bigrams stored by the program. See Appendix Exhibits 1-12 for analysis.

For further details about the top 100 bigrams, please refer to the attached file: Collocation Analysis.xlsx. The worksheets in the workbook are named by domain name (biology or politics) and type of test (frequency, T-Score, or χ^2-Score).

Odd Couples: Headings & Textbook conventions

A few of the unexpected collocation candidates relate to common headings, or textbook conventions that occur in some of the texts. These bigrams were highlighted in red. For example, "WHY I BELIEVE IN FAIR TAXATION OF CHURCH PROPERTY" is a recurring heading in one of the political corpora, which

explains FAIR TAXATION and CHURCH PROPERTY as common collocations. Similarly, in the biology tests, 'Diagram Sheet', 'See text' and 'compare Section', are commonly used for referring to other areas of the textbook. Fixes to such problems are discussed in the 'Areas for Improvement' section.

T-Score Analysis

In both domains, the top ranked T-Score rankings did not vary compared to Frequency rankings. Referring to Exhibits 2 and 8, there did not exist an absolute difference greater than 1 for any instance.

On the other hand, per Exhibits 5 and 11, bottom ranked bigrams, were slightly different. In the biology tests, the major downgrades in ranking related to bigrams 'layer cells', 'branchial arch', 'dorsal side' and "compare Section". Justification could be related to the fact that words like 'layer', 'cells', 'arch', 'side' and 'compare' are relatively promiscuous and could be used alongside other alternative words. For example, some instances of words following 'branchial' include 'arch', 'arches', 'becomes', 'arteries', 'vessel', 'is', 'and', ')', etc. In contrast, with a term like "vas deferens", a search through the input file shows that 'deferens' almost always follows 'vas'.

In the political tests, rank downgrades occur with bigrams 'good reason', 'go home', 'great many', 'most men', and 'go out'. This appears reasonable considering none of these expressions tend to establish an idea significantly larger than the sum of the parts. Additionally, upgraded bigrams include political terms like 'public education', 'third world' and 'trade deficit'.

Therefore, it would appear that the T-Score method slightly improved the results when compared to the frequency method. This is especially relevant in the middle-to-lower ranks on the lists.

χ^2-Score Analysis

As seen in Exhibit 3, eight of the top ten biology terms were upgraded using χ^2 analysis (highlighted in blue). To approximately assess the terminological dependence of each bigram, we can type the first

word into Google search to see if autocomplete tends to show the second word. For seven of these eight instances, this was successful, which somewhat shows the strength of the χ^2 method.

For the bottom 10 biology bigrams, Exhibit 6 shows a few downgrades identical to the T-Rankings. Additionally, it has downgraded 'young rabbit', 'body wall' and 'brain case', which at first glance appear to be poor candidates for collocation, since the words are not highly technical.

In regards to politics, χ^2 approach appears to be an improvement also. It upgrades bigrams like 'Warsaw Pact', 'carbon dioxide', 'Saddam Hussein', 'Clinton Administration', and 'fossil fuels' while downgrading 'church property', 'New America', 'been done', 'other people', 'young women', 'American people', and more. Please refer to 'Areas of Improvement' for an explanation about the bigram 'Miss K'.

In conclusion, the χ^2 method appears to be the most affective of the three methods.

Areas of Improvement

More refined POS tag set

As stated early, certain content-lacking words that commonly appeared were used to filter out certain bigram options. Instead of a manual fix like this, one might use a more refined POS tag set, which classifies such words together, while keeping them decoupled from content words of the same POS.

Updated text corpora

As seen in the Biology experimentation, there were certain bigrams that did not relate to text content. Instead such text was commonly repeated for inconvenient purposes. For example, the bigram 'Diagram Sheet' occurred 24 times. But, such a bigram was not common because it was a collocation. Instead, it frequently occurred as a formatting convention in the textbook used for the corpus. Another example occurred, when the address of an article's supplier was written on every page of the article. Its

components often appeared as a common bigrams until they were removed from the corpus. Therefore, if corpora could be updated to exclude such distractions, it may likely improve collocation identification.

Larger sample size

Adding more domain-specific text to the corpora could help filter away context-specific bigrams. For example, in the politics results, Miss K, Mrs. Garner and Jane Alice were all common bigrams. The only reason for this was because they were people of discussion in one of the larger political corpora. With more articles (likely not about the above stated individuals), their relative frequency and scores would decrease.

More comprehensive data analysis

In our testing, only a subset of the top 100 bigrams were analyzed. Increasing such numbers could help identify more problem areas in collocation identification.

Experimentation

Experimentation regarding case-sensitivity and variation in word sense or POS could be performed to assess whether our definition of a word is necessary.

Program Execution Instructions

1. Unzip project file.
2. Open Eclipse.
3. Import *CollocationSearch* as existing project into workspace.
4. Run Driver class as Java Application. Program will commence in Console window.
5. Enter desired file's number. Press Enter.
6. Once program has finished running, see related output file (by domain name) in outputFiles directory. (Open in Microsoft Excel for better column alignment and data manipulation.)

Appendix

Exhibit I: Biology – Frequency Top Ten

Number	Word 1	Word 2	Frequency	TScore	ChiScore
1	alimentary	canal	28	5.287566461	29539.1325
2	dorsal	aorta	27	5.185200602	11690.27078
3	spinal	cord	27	5.193225946	35538.77274
4	Diagram	Sheet	24	4.889734253	11620.64802
5	body	cavity	22	4.662643972	3584.924821
6	portal	vein	21	4.575010828	11624.17103
7	vena	cava	21	4.581276048	48129.19808
8	middle	line	19	4.354012404	15070.66777
9	vertebral	column	19	4.356748489	30077.27247
10	nervous	system	18	4.237711119	13905.51293

Exhibit 2: Biology – T-Score Top Ten

Word 1	Word 2	Frequency	TScore	ChiScore	T Rank	Freq Rank	Difference
alimentary	canal	28	5.287566461	29539.1325	1	1	0
spinal	cord	27	5.193225946	35538.77274	2	3	-1
dorsal	aorta	27	5.185200602	11690.27078	3	2	1
Diagram	Sheet	24	4.889734253	11620.64802	4	4	0
body	cavity	22	4.662643972	3584.924821	5	5	0
vena	cava	21	4.581276048	48129.19808	6	7	-1
portal	vein	21	4.575010828	11624.17103	7	6	1
vertebral	column	19	4.356748489	30077.27247	8	9	-1
middle	line	19	4.354012404	15070.66777	9	8	1
nervous	system	18	4.237711119	13905.51293	10	10	0

Exhibit 3: Biology – χ^2-Score Top Ten

Word 1	Word 2	Frequency	TScore	ChiScore	Chi Rank	Freq Rank	Difference
vas	deferens	6	2.449382877	68765	1	95	-94
sacculus	rotundus	6	2.449347254	58940.57135	2	88	-86
vasa	efferentia	6	2.449347254	58940.57135	3	96	-93
corpus	callosum	6	2.449305694	50519.51004	4	69	-65
carbon	dioxide	13	3.604948256	49659.99932	5	15	-10
vena	cava	21	4.581276048	48129.19808	6	7	-1
truncus	arteriosus	9	2.999609763	46029.4954	7	36	-29
bilateral	symmetry	6	2.449276009	45841.33316	8	64	-56
foramen	lacerum	6	2.449276009	45841.33316	9	75	-66
spinal	cord	27	5.193225946	35538.77274	10	3	7

Exhibit 4: Biology – Frequency Bottom Ten

Number	Word 1	Word 2	Frequency	TScore	ChiScore
91	sympathetic	chain	6	2.448314195	11453.8329
92	third	ventricle	6	2.442115841	1954.750372
93	transverse	process	6	2.444680677	2980.351253
94	TRUE	vertebrata	6	2.445701862	3764.818738
95	vas	deferens	6	2.449382877	68765
96	vasa	efferentia	6	2.449347254	58940.57135
97	ventral	wall	6	2.425776885	606.211321
98	See	text	5	2.234978602	9544.444245
99	abdominal	vein	5	2.225866859	1079.114428
100	anterior	part	5	2.227024525	1216.8272

Exhibit 5: Biology – T-Score Bottom Ten

Word 1	Word 2	Frequency	TScore	ChiScore	T Rank	Freq Rank	Difference
third	ventricle	6	2.442115841	1954.750372	91	92	-1
branchial	arch	6	2.43908791	1388.432386	92	66	26
layer	cells	6	2.438624814	1329.712015	93	80	13
ventral	wall	6	2.425776885	606.211321	94	97	-3
side	view	6	2.417886452	452.6145952	95	90	5
dorsal	side	6	2.390706313	238.4127198	96	72	24
compare	Section	6	2.373043379	181.0977559	97	68	29
See	text	5	2.234978602	9544.444245	98	98	0
anterior	part	5	2.227024525	1216.8272	99	100	-1
abdominal	vein	5	2.225866859	1079.114428	100	99	1

Exhibit 6: Biology – χ^2-Score Bottom Ten

Word 1	Word 2	Frequency	TScore	ChiScore	Chi Rank	Freq Rank	Difference
brain	case	8	2.811963175	1347.140079	91	39	52
layer	cells	6	2.438624814	1329.712015	92	80	12
anterior	part	5	2.227024525	1216.8272	93	100	-7
abdominal	vein	5	2.225866859	1079.114428	94	99	-5
body	wall	11	3.279319297	951.9294912	95	26	69
ventral	wall	6	2.425776885	606.211321	96	97	-1
young	rabbit	7	2.613048426	551.997536	97	59	38
side	view	6	2.417886452	452.6145952	98	90	8
dorsal	side	6	2.390706313	238.4127198	99	72	27
compare	Section	6	2.373043379	181.0977559	100	68	32

Exhibit 7: Politics – Frequency Top Ten

Number	Word 1	Word 2	Frequency	TScore	ChiScore
1	United	States	41	6.40180112	124884.9307
2	church	property	27	5.185306988	12421.58945
3	Miss	K	23	4.795385418	142622.3371
4	CHURCH	PROPERTY	22	4.68999822	142581.1534
5	FAIR	TAXATION	22	4.690109567	168509
6	long	term	21	4.580540496	41448.66536
7	American	Mandate	19	4.358368638	106715.3992
8	Persian	Gulf	19	4.358627327	160082.5999
9	George	Bush	18	4.24216789	109186.9912
10	New	American	18	4.240936923	39544.82416

Exhibit 8: Politics – T-Score Top Ten

Word 1	Word 2	Frequency	TScore	ChiScore	T Rank	Freq. Rank	Difference
United	States	41	6.40180112	124884.9307	1	1	0
church	property	27	5.185306988	12421.58945	2	2	0
Miss	K	23	4.795385418	142622.3371	3	3	0
FAIR	TAXATION	22	4.690109567	168509	4	5	-1
CHURCH	PROPERTY	22	4.68999822	142581.1534	5	4	1
long	term	21	4.580540496	41448.66536	6	6	0
Persian	Gulf	19	4.358627327	160082.5999	7	8	-1
American	Mandate	19	4.358368638	106715.3992	8	7	1
George	Bush	18	4.24216789	109186.9912	9	9	0
New	American	18	4.240936923	39544.82416	10	10	0

Exhibit 9: Politics – χ^2-Score Top Ten

Word 1	Word 2	Frequency	TScore	ChiScore	Chi Rank	Freq. Rank	Diff
FAIR	TAXATION	22	4.690109567	168509	1	5	-4
Warsaw	Pact	7	2.645696357	168509	2	77	-75
carbon	dioxide	7	2.645696357	168509	3	78	-75
Saddam	Hussein	11	3.316516537	168509	4	34	-30
Clinton	Administration	10	3.162183828	168509	5	40	-35
Persian	Gulf	19	4.358627327	160082.5999	6	8	-2
fossil	fuels	9	2.999902081	151657.2	7	52	-45
Soviet	Union	15	3.872764994	148682.6469	8	16	-8
Miss	K	23	4.795385418	142622.3371	9	3	6
CHURCH	PROPERTY	22	4.68999822	142581.1534	10	4	6

Exhibit 10: Politics – Frequency Bottom Ten

Number	Word 1	Word 2	Frequency	TScore	ChiScore
91	public	education	7	2.631150247	1251.026256
92	third	world	7	2.637731324	2283.070704
93	trade	deficit	7	2.643899683	9703.911312
94	wanted	children	7	2.600667342	397.1033617
95	whole	story	7	2.635775394	1833.892378
96	year	old	7	2.633853109	1536.798206
97	young	doctor	7	2.600443038	395.0646319
98	Mrs.	Garner	6	2.448733844	18375.49079
99	New	Hampshire	6	2.448864673	21974.21721
100	economic	power	6	2.393124412	249.152547

Exhibit 11: Politics – T-Score Bottom Ten

Word 1	Word 2	Frequency	TScore	ChiScore	T Rank	Freq. Rank	Difference
good	reason	7	2.618216882	658.4041413	91	83	8
go	home	7	2.613968565	568.6312134	92	81	11
great	many	7	2.611353181	524.423943	93	84	9
most	men	7	2.610274279	508.2069568	94	87	7
wanted	children	7	2.600667342	397.1033617	95	94	1
young	doctor	7	2.600443038	395.0646319	96	97	-1
go	out	7	2.584840456	290.5190621	97	82	15
New	Hampshire	6	2.448864673	21974.21721	98	99	-1
Mrs.	Garner	6	2.448733844	18375.49079	99	98	1
economic	power	6	2.393124412	249.152547	100	100	0

Exhibit 12: Politics – χ^2-Score Bottom Ten

Word 1	Word 2	Frequency	TScore	ChiScore	Chi Rank	Freq. Rank	Diff
young	women	8	2.78901496	558.1820868	91	73	18
great	many	7	2.611353181	524.423943	92	84	8
most	men	7	2.610274279	508.2069568	93	87	6
American	people	9	2.941387323	443.1913101	94	45	49
been	done	8	2.777558912	429.4753608	95	64	31
wanted	children	7	2.600667342	397.1033617	96	94	2
young	doctor	7	2.600443038	395.0646319	97	97	0
go	out	7	2.584840456	290.5190621	98	82	16
economic	power	6	2.393124412	249.152547	99	100	-1
other	people	8	2.705612407	168.9775284	100	69	31

References

Political Text Files

http://www.gutenberg.org/cache/epub/7370/pg7370.txt

http://textfiles.com/politics/court.txt

http://textfiles.com/politics/economy.txt

http://textfiles.com/politics/tsongtxt.txt

http://textfiles.com/politics/taxch.d

http://textfiles.com/politics/share02.txt

http://textfiles.com/politics/rights.txt

http://textfiles.com/politics/abortion.txt

Biology Text File

http://www.gutenberg.org/files/21781/21781-h/21781-h.htm

Philosophy Text File

http://www.gutenberg.org/cache/epub/16406/pg16406.html

Programming Tools & Libraries

Stanford POS Tagger

http://nlp.stanford.edu/software/tagger.shtml

Apache Commons Mathematics Library – χ^2 Score Calculation

http://commons.apache.org/proper/commons-math/