Statistical NLP Approach for Collocation Extraction from Newspaper Text

GurinderPal Singh Gosal

Department of Computer Science,
Punjabi University, Patiala

Abstract–Statistical NLP is an approach of doing natural language processing to resolve the problems usually encountered with traditional NLP. The task of finding some interesting combination of wordsfrom text in large corpora, known as collocation extraction, is one of many tasks in statistical NLP. Collocations have multiple applications and the methods of collocation extraction are influenced by their intended use. In the present task,the effectiveness of selecting bigram collocations from textis analysed and evaluatedby applying different statistical NLP approaches ranging from just raw frequency count to the usage of more sophisticated statistical association measures. It is observed that the collocations extracted by filtering bigrams using POS taggers seem to give the best results. It is also interesting to observe that some bad collocations are selected and verified by all approaches.

Keywords— Collocation, n-gram, POS-filtering, association measure, t-test, chi-square test, bigram

1. INTRODUCTION

Natural Language Processing (NLP) is the research field aimed at exploiting rich knowledge resources with the goal of understanding, extraction and retrieval from unstructured text. One of the many definitions of Natural Language Processing is "a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications"[1]. Research in natural language processing has been going on for several decades dating back to the late 1940s.

There are some well documented problems in the literaturewhich are encountered with traditional NLP approaches. For example, there are a large number of analyses producedwhich are highly ambiguous when we try to process longer sentences with realistic grammars. To overcome these problems with traditional NLP, statistical NLP is often used. In statistical NLP, probabilistic, stochastic and statistical methods are used to perform different NLP tasks and it is usually carried out with the use of several machine learning algorithms. Learning algorithms can be generally classified into three types: supervised learning, semi-supervised learning and unsupervised learning.Supervised learning technique is based on the idea of studyingthe features of positive and negative examples over a large collection of annotated corpus. Semi-supervised learning uses both labeled data and unlabeled data for the learning process to reduce the dependence on training data. In the unsupervised learning, decisions are made on the basis of unlabeled data. The methods of unsupervised learning are mostly built upon clustering techniques, similarity based functions and distribution statistics.

One of the fundamental tasks in the processing of natural language is text mining. Most of the text mining system depends upon the methods and tools of NLP. Text mining can be defined as a knowledge extraction method to detect useful and previously unknown information from a document or set of texts by identifying facts inherent and inexplicit in the data [2].There are many

tasks that are performed under the umbrella of text mining, for example, entity/relation extraction, event extraction, text categorization, clustering, sentiment analysis/opinion mining, taxonomies generation and document summarization.

The task of finding some interesting combination of words in text of large corpora, usually referred to as collocations,is one of many tasks performed in statistical NLP. There has been some lack of consensus as far as defining the concept of collocations is concerned. A collocation,in its simplest form, can be viewed as an expression consisting of two or more words. Another related term to the collocation is the word *n*-gram[1], which represents any sequence of *n* words. However, for acombination of words to qualify as a collocation, it should possess certain characteristics. Collocation expression forms a semantic unit and the whole has an independent existence beyond the individual parts, characteristic often referred to as non-compositionality. Another important characteristic of expression being a collocation is non-substitutability that means we cannotreplace a word in a collocation with some other word having similar or even same meaning. In another characteristic of non-modifiability, we find that it is not feasible to alter collocation if we try grammatical transformations or put some extralexical material. One example of such a type is idioms.

Collocations find their application and use in multiple tasks and that also derive what type of method or approach will be used to extract collocations. These tasks making use of collocations include, for example,automatic language generation, word sense disambiguation in multilingual lexicography, finding multipleword combinations in text for indexing purposes in information retrieval, improving text categorization systems etc. [3]. The strength of association of two (or more) words is often represented in terms of "Association Measures (AMs)". Association measures are usually defined for bigrams and extended to higher associations. Choice of association measures is very important because it can derive the process of collocation extraction. There are many categories of measuring association that are used by researchers for collocations, such as, sorting by pure frequencies, measures based on hypothesis testing etc. In one of the basic measure of sorting by pure frequencies, scoring is based on the occurrence or frequencies in the corpus. In the testing of hypothesis measure, the *null hypothesis* of no association between the words beyond chance occurrences is tested. Pearson's chi-square*, t* -test, likelihood ratios are the commonly used hypothesis testing measures for collocations.

The aim of the experiments done in this workis to analyse and evaluate the effectiveness of selecting collocations by applying different statistical NLP approaches ranging from just raw frequency count to the usage of more sophisticated statistical association measures. Collocations are comprised of variable length n-grams, however, in the experiments done for this task the bigrams (digrams), the collocations made of two elements, are used.The experiments and results are detailed and evaluated in subsequent sections.

2. EXPERIMENTATION

The following is the experimental setup for the task:

2.1. Corpora

Source: The 20 Newsgroups data setis a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups [4]. The organization of these data is done into 20 different newsgroups. Each newsgroup corresponds to a different topic. Some of the newsgroups are very closely related to each other, while others are highly unrelated.

[1]A word *n*-gram consisting of two words is called a digram, word *n*-gram consisting of three words is called a trigram and a word *n*-gram consisting of four words is called a tetragram.

Coverage: Our reference corpus is from the version of 20 Newsgroups, sorted by date into training(60%) and test(40%) sets that does not include cross-posts (duplicates) and does not include newsgroup-identifying headers (Xref, Newsgroups, Path, Followup-To, Date). Further our subset covers two topics or domains of ***sci.space*** *(Space*) and ***rec.sport.hockey*** *(Hockey*).

Size: In our experimentation, size of corpus is 125444 words with 22470 word types. The raw bigrams in the corpus are 125443 while the number of distinct bigram types is 78990.

Word: For our purposes the word is considered as a word token without punctuation marks.

2.2. Software:

The collocation generator tool is implemented in Java language.The developedcode performs certain functions implemented in terms of several routines that include: getting text from corpora, extracting words as tokens in raw from, refining the words with some primitive rules, extracting bigrams (2 word pairs), frequency engine for frequency count of words and bigrams, filtering engine for applying POS tag filtering on bigrams, implementing statistical tests of "t" and "chi-square" for hypothesis testing for collocations.

2.3. POS Filtering Tagger:

Stanford Log-linear Part-Of-Speech Tagger [5,6] was used for POS filtering ofbigrams. The system requires Java 1.6+ to be installed. You'll need somewhere between 60 and 200 MB of memory to run a trained tagger (i.e., you may need to give java an option like java -mx200m) depending on whether you're running 32 or 64 bit and the complexity of the tagger model.

In this application, the *english-left3words-distsim.tagger model*has been used. It's nearly as accurate (96.97% accuracy vs. 97.32% on the standard WSJ22-24 test set) and is an order of magnitude faster.The English taggers use the Penn Treebank tag set [7].

3. RESULTS AND DISCUSSION

3.1 Base Measure of Frequency Count

The base measure of finding collocations in the text is frequency count. The co-existence of two words is significant if it occurs many times. However this approach has obvious drawback of having a large false positives. The list of the 50 most frequent bigrams in the corpus along with their frequency is shown in *Table*3.1.

Note that the first interesting collocations in the domains, "Hockey League"and "Power play", appear only at the order 35 and 39 not in the list of first 10.

Table 3.1: The 50 most frequent bigrams in the corpus along with their frequency

Sr. No.	Bigram	Frequency
1	of the	601
2	in the	434
3	on the	237
4	for the	224
5	Subject Re	217
6	to the	201
7	In article	139
8	and the	127
9	will be	120
10	from the	118
11	University of	115
12	to be	115
13	is a	103
14	with the	102
15	for a	97
16	that the	96
17	Organization University	88
18	I think	81
19	the first	78
20	in a	77
21	at the	77
22	by the	73
23	is the	72
24	of a	70
25	have been	70
26	would be	67
27	have to	64
28	it is	55
29	going to	55
30	I don't	54
31	one of	53
32	the NHL	53
33	out of	52
34	to get	50
35	Hockey League	49
36	have a	47
37	the same	46
38	be a	46
39	Power play	46
40	with a	46
41	the team	45
42	as a	45
43	would have	45
44	the game	44

45	is not	43
46	into the	43
47	and other	42
48	was a	42
49	it was	42
50	and a	42

3.2 Filtering Collocations using POS Filtering

The frequency count was a very simple measurebut it is interesting to see if some kind of filtering, such as, based on part-of-speechimproves the result or not. Justeson and Kartz[8] had suggested some POS tag patterns that can be used for filtering collocations effectively.

The tag patterns as shown in the *Table3.2*were used to improve the filtering in our case.

Sr. No.	Tag Pattern
1	Adjective Noun (JJ/JJR/JJS NN/NNP/NNPS/NNS)
2	Noun Noun (NN/NNP NN/NNP/NNPS/NNS)

Table 3.2: Tag Patterns used for POS filtering of bigrams

The Stanford POS tagger [5, 6] used in the extractingbigrams using the above POS tag patterns improves the result considerably. The table 3.3shows the 50 top ranked collocations after POS filtering.

Sr. No.	Filtered Bigram	Frequency
1	Subject Re	217
2	Organization University	88
3	Hockey League	49
4	Power play	46
5	New York	39
6	Distribution world	39
7	space station	34
8	Los Angeles	33
9	St Louis	31
10	San Jose	27
11	anonymous FTP	27
12	Tampa Bay	27
13	New Jersey	25
14	First period	23
15	Third period	23
16	Roger Maynard	23
17	Second period	23
18	Ron Baalke	22
19	Jet Propulsion	21
20	Many Europeans	20
21	Space Shuttle	20
22	Stanley Cup	20

23	North Carolina	20
24	power play	19
25	Air Force	19
26	last year	18
27	Maple Leafs	18
28	Red Wings	17
29	Research Center	17
30	International Space	17
31	Jon Leech	17
32	Subject Space	17
33	Inc Lines	17
Sr. No.	**Filtered Bigram**	**Frequency**
34	Carolina Chapel	16
35	draft pick	16
36	Henry Spencer	16
37	Chapel Hill	16
38	Calgary Flames	16
39	Space Digest	16
40	Mary Shafer	16
41	Space Flight	16
42	j mccall	16
43	Flight Center	15
44	Winnipeg Jets	15
45	Propulsion Laboratory	15
46	Montreal Canadiens	15
47	Space Station	15
48	NY Rangers	15
49	Space FAQ	15
50	Deepak Chhabra	15

Table 3.3: The best 50 collocations found by applying filtering of POS taggers using Stanford tagger.

Clearly certain collocations representing both the domains of our corpus i.e. “Space” and “Hockey” appear in the top 10 of POS filtered collocations, such as, “Hockey League” (Hockey) and “space station”(Space). Further it is noticeable that “Hockey League” and “Power play” that appear as lowerasplaces 35 and 39 in frequency count measure T*able3.1*, are now found at place 3 and 4 (shaded) in POS filtered *Table3.3*.

3.3. Association Measure by t-Test:

The above two methods to find collocations are based onthe frequency and it is commonsensical to check whether results are merely accidental. Few statistical association measures, such as, t-test and chi-square, help in exploring how probable or improbable it is that a certain collocation will occur [9, 10]. The critical value used for *t is **2.576*** at confidence level of **α=0.005.** If value of t is larger than critical threshold, we can reject the null hypothesis(that bigram components occur independently) with **99.5%** confidence.

In t-test conducted on all bigrams (not considering the POS filtered bigrams), we see that the collocations in the upper part of *Table 4*, such as, “toallow”, “we do”, “let the”etc., are

compositional by nature and cannotbe elevated to status of collocations although frequency is more than 7. While in the lower part of *Table 4* the collocations are accepted because null hypothesis is rejected by value of t. Some of the collocations accepted by t-test obviously do not qualify as collocation.

Now if we use t-test to consider filtered bigrams by POS and try to find out whether high frequencycollocations are accidental or not, the results are shown in *Table 5*. It is observed that the most filtered bigrams areoccurring significantly more than by chance and for most of the cases null hypothesis of independence is rejected.

Bigram	Prob. Word1	Prob. Word2	Mean Distribution	Mean Sample	Frequency	t-value	Remark
to allow	0.017777	8.77E-05	1.56E-06	5.58E-05	7	2.571853	Smaller than threshold (t < 2.576)
we do	0.001172	0.001347	1.58E-06	5.58E-05	7	2.57091	Smaller than threshold (t < 2.576)
No not	5.82E-04	0.002718	1.58E-06	5.58E-05	7	2.570759	Smaller than threshold (t < 2.576)
let the	1.51E-04	0.038416	5.82E-06	6.38E-05	8	2.570382	Smaller than threshold (t < 2.576)
Subject Re	0.002694	0.001841	4.96E-06	0.00173	217	14.68873	Greater than threshold (t > 2.576)
Hockey League	7.65E-04	5.02E-04	3.84E-07	3.91E-04	49	6.99314	Greater than threshold (t > 2.576)
have to	0.003906	0.017777	6.94E-05	5.10E-04	64	6.911204	Greater than threshold (t > 2.576)
by the	0.003125	0.038416	1.20E-04	5.82E-04	73	6.781535	Greater than threshold (t > 2.576)
Power play	5.42E-04	0.001483	8.04E-07	3.67E-04	46	6.767491	Greater than threshold (t > 2.576)

Table 4: The collocations found by applying t-test on bigrams (not filtered).

Bigram	Prob. Word1	Prob. Word2	Mean Distribution	Mean Sample	Frequency	t-value	Remark
Subject Re	0.00269443	0.001841	4.96E-06	0.008867	217	33.33252	(t > 2.576)
Organization University	0.00267849	0.001427	3.82E-06	0.003596	88	21.21587	(t > 2.576)
Hockey League	7.65E-04	5.02E-04	3.84E-07	0.002002	49	15.84514	(t > 2.576)
Power play	5.42E-04	0.001483	8.04E-07	0.00188	46	15.3488	(t > 2.576)
New York	8.29E-04	3.43E-04	2.84E-07	0.001594	39	14.13631	(t > 2.576)
Distribution world	6.62E-04	6.70E-04	4.43E-07	0.001594	39	14.1349	(t > 2.576)
space station	0.00164217	3.67E-04	6.02E-07	0.001389	34	13.1957	(t > 2.576)
Los Angeles	2.87E-04	2.95E-04	8.46E-08	0.001348	33	13.00502	(t > 2.576)
St Louis	3.91E-04	2.79E-04	1.09E-07	0.001267	31	12.60448	(t > 2.576)
Tampa Bay	2.47E-04	2.47E-04	6.11E-08	0.001103	27	11.76357	(t > 2.576)

Table 5: The collocations found by applying t-test on bigrams that are filtered by POS taggers using Stanford tagger

3.4 Association Measure by chi-squareTest

The critical value used for chi-square is 3.841 (the statistic has one degree of freedom for a 2 X 2 table) at confidence level of α=0.005. If value of t is larger than critical threshold, we can reject the null hypothesis (that bigram components occur independently) with 99.5% confidence.

Using the Chi-square test on POS filtered bigrams again show that most bigrams do not occur by chance. If we look at the *table 6*, we see that the results are similar to t-test.

Bigram	cell_O 11	cell_O 12	cell_O 21	cell_O 22	N	X2	Remark
Subject Re	217	2	95	24159	24473	16797.17	Greater than threshold (X2 > 3.841)
Organization University	88	82	225	24078	24473	3455.787	Greater than threshold (X2 > 3.841)
Hockey League	49	13	40	24371	24473	10616.36	Greater than threshold (X2 > 3.841)
Power play	46	56	18	24353	24473	7894.46	Greater than threshold (X2 > 3.841)
Distribution world	39	15	6	24413	24473	15302.22	Greater than threshold (X2 > 3.841)
New York	39	1	62	24371	24473	9188.765	Greater than threshold (X2 > 3.841)
space station	34	7	118	24314	24473	4507.354	Greater than threshold (X2 > 3.841)
Los Angeles	33	3	2	24435	24473	21147.02	Greater than threshold (X2 > 3.841)
St Louis	31	4	17	24421	24473	13984.6	Greater than threshold (X2 > 3.841)
Tampa Bay	27	4	3	24439	24473	19177.48	Greater than threshold (X2 > 3.841)

Table 6: The collocations found by applying Chi-Square test on bigrams that are filtered by POS taggers using Stanford tagger

The t-test has the problem of underlying assumption of normality and chi-square has problem with numbers in (2X2)table being small. In our case the problem of numbers of being small in (2X 2) table is very frequentfor majority of bigrams.

4. CONCLUSIONS AND FUTURE WORK

In the present set of experiments it is seen that the collocations extracted by filtering bigrams using POS taggers seem to give the best results. The methods work well on both types of corpora. We see collocations, such as, Jet Propulsion, Space Shuttle, Space Digest, Space Flight, are clearly the collocations which represent the “Space” domain are selected while the collocations Hockey League, Power play, Stanley Cup, Red Wings are selected for domain “Hockey”.

“Subject Re” is appearing as one of the top collocations in various measures and even accepted by t-test and chi-square test, although it appears to not qualify as a collocation.NEXT FAQ, Inc Lines etc. are another examples of bad collocations selected and verified by different approaches.

In future experimentation, it will be interesting to observe the effects of stemming and elimination of stop words on the results of collocation extraction. Alsothe effect ofapplying the same set of tests on corpora much larger in size than used in present experimental setup can be explored.

REFERENCES

1. Liddy, E. D. (2001). Natural language processing. In: Encyclopedia of Library and Information Science, 2nd edn. Marcel Decker, Inc., New York.
2. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., &Verspoor, K. (2014). Biomedical text mining: State-of-the-art, open problems and future challenges. In Interactive Knowledge Discovery and Data Mining in Biomedical Informatics (pp. 271-300). Springer Berlin Heidelberg.
3. Petrović, S. (2007). Collocation extraction measures for text mining applications (Doctoral dissertation, Fakultetelektrotehnike i računarstva, Sveučilište u Zagrebu).
4. The 20 Newsgroups data set, available at http://qwone.com/~jason/20Newsgroups/,2007.
5. Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.
6. Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.
7. Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz: Building a Large Annotated Corpus of English: The Penn Treebank, in Computational Linguistics, Volume 19, Number 2 (June 1993), pp. 313—330
8. J. Justeson, S. Katz (1995). Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text
9. Foundations of Statistical Natural Language Processing, by Chris Manning and HinrichSchutze, MIT Press, 1999.
10. Speech and Language Processing, Daniel Jurafsky& James H. Martin. Prentice Hall, 2nd edition, 2008.