Automatic Suggestion of Significant Terms for a Predefined Topic

Joe Zhou and Pete Dapkus

LEXIS-NEXIS, a Division of Reed Elsevier, Inc.
9555 Springboro Pike
Miamisburg, OH 45342
{joez,peted}@lexis-nexis.com

ABSTRACT

This paper presents a preliminary experiment in automatically suggesting significant terms for a predefined topic. The general method is to compare a topically focused sample created around the predefined topic with a larger and more general base sample. A set of statistical measures are used to identify significant word units in both samples. Identification of single word terms is based on the notion of word intervals. Two-word terms are identified through the computation of mutual information, and the extension of mutual information assists in capturing multi-word terms. Once significant terms of all these three types are identified, a comparison algorithm is applied to differentiate terms across the two data samples. If significant changes in the values of certain statistical variables are detected, associated terms will selected as being topic-oriented and included in a suggested list. To check the quality of the suggested terms, we compare them against terms manually determined by the domain expert. Though overlaps vary, we find that the automatical suggestion provides more terms that are useful for describing the predefined topic.

1. INTRODUCTION

As we are facing the growing amount of on-line text, the use of text analysis techniques to access information from electronic sources has become more popular and, at the same time, more difficult. Currently, the effectiveness of such techniques is evaluated not only on how easily they can be applied to text sources to extract information and represent it in a systematic format (Walker 1983), but also on whether they can be applied to large text corpora of several tens of thousand of words.

One of the applications of text analysis is to identify and extract significant terminology from running text. Choueka (1988), for example, describes an experiment for locating interesting collocational expressions from large textual databases. A collocational expression, as Choueka defines it, is "sequences of words whose unambiguous meaning cannot be derived from that of their components". Other representative collocation research can be found in Church and Hanks (1990) and Smadja (1993). Though all statistically-based, their definitions of collocations are different from one another. Unlike Choueka (1988), Church and Hanks (1990) identify as collocations both interrupted and uninterrupted sequences of words. Unlike Church and Hanks (1990), Smadja (1993) goes beyond the "two-word" limitation and deals with "collocations of arbitrary length".

The primary goal of collocation research is to build a comprehensive lexicographic toolkit, or to assist automatic language generation applications. Therefore, the focus is on the extraction of all interesting word patterns without distinction of domain specificity. Identifying domain-specific terminology is another research effort. Gierl and Frost (1992) describe their approach to extracting terminological knowledge from medical texts. Following Church and Hanks (1990), they use mutual information to select significant two-word patterns, but, at the same time, a lexical inductive process is incorporated which, as they claim, can improve the collection of domain-specific terms. Justeson and Katz (1993) introduce an algorithm by which technical terms in running text can be identified. Prior to the development of their algorithm, they performed a thorough study on the linguistic properties of technical terminology. They report that, structurally, technical terms make heavy use of noun compounds. In technical terminology, word constituents are limited to adjectives, nouns and occasionally prepositions. Verbs, adverbs, or conjunctions are extremely rare. At the discourse level, technical terms tend to be repetitive. With these observations in mind, they developed an algorithm which has proved to be effective and domain independent.

In this paper, a preliminary experiment is presented in automatically suggesting significant terms for a predefined topic. The general method is to compare a topic focused sample based on the predefined topic with a larger and more general base sample. A set of statistical measures are used to identify significant word units in both samples. Identification of single word terms is based on the notion of word intervals. Two-word terms are identified through the computation of mutual information, and an extension of mutual information assists in capturing multi-word terms. Once significant terms of all these three types are identified, a comparison algorithm is applied to differentiate terms across the two samples. If significant changes in the values of certain statistical variables are detected, associated terms are selected from the focused sample as being topic-oriented and included in a suggested list.

To check the quality of the suggested terms, we compare them against terms manually determined by a domain expert. Though the numbers of matches vary, we find that our automatic suggestion process provides more terms (than the manual process) that are useful for describing the predefined topic.

2. METHODOLOGY

2.1 Manual versus Automatic Term Suggestion

To manually select significant terms for a predefined topic, the domain expert first creates a topic focused sample from one specific source or a combination of sources. Then, he or she reads the documents, providing a relevance judgment (i.e. a reader-assigned score) to each document. By carefully examining relevant documents in the focused sample, a list of terms that are deemed to be significant for the definition of the topic is identified. In many cases, it is possible that the domain expert would introduce some terms based on his or her own professional knowledge about the topic. These terms may be highly prominent for the topic, yet may not necessarily occur in the focused sample.

For automatic suggestion of topical terms, initial attempts were made using the sample documents the domain expert created. The results were not impressive. The statistical information generated from the sample documents was not rich and sufficient enough for any discriminative judgment. Our experience showed that, to draw terms that are reflective of a given topic, a much larger and more general base sample is required. Such a base sample should be randomly sampled from the same source as the focused sample and it should contain an array of different topics. Once the baseline statistics are generated from both data collections, a meaningful comparison could spot terms that occur with unusual frequency in the focused sample. These terms would constitute good candidates for topically sensitive terminological units (Steier and Belew 1994).

2.2 Focused Sample and Base Sample

For our experiments of automatic term suggestion, we selected a predefined topic called "European Politics and Business". The focused sample was originally created by the domain expert using the 1988 United Press International (UPI). Table 1 presents statistical information about this dataset. After reading each of the relevant documents found in the focused sample, the domain expert manually determined 347 topical terms. Table 2 provides the statistical breakdown of these terms.

Table 1: Focused and Base Samples

Data File	Source/Name	Size (bytes)	Unique Words
Focused Sample	Sample from 1988 UPI	1,015200	12,065 (5,045*)
Base Sample	Sample from 1987,1988,1989 UPI	27,322,598	73,583 (33,114*)

*** only words which occur more than 3 times were used in the experiments**

Table 2: Predefined Topic and its Manually Determined Topical Terms

Predefined Topic	one-word	two-word	multi-word	total terms
European Politics & Business	276	36	35	347

Since the focused sample was drawn from the source of 1988 UPI, the construction of its corresponding base sample was also initiated from the same source of the same year. Our experiments demonstrated that, in order to obtain a random assortment of topics to be included in the base sample, it may be meaningful to sample documents from the time period before and after the focused documents. Therefore, the final base sample was created by randomly drawing documents from the years of 1987, 1988 and 1989. The size of this dataset is about 27 times larger than the sample data file (see Table 1).

Though the ratio between the focused and base samples was arbitrary, in order to generate meaningful statistics, we felt that the base sample should be at least 20 times larger in size than the focused sample. (For the sake of discussion, hereafter, we may sometimes refer to the focused sample as "focused" and the base sample as "base".)

2.3 Experimental Procedure

The general method we adopted is as follows. First, we identified statistically significant terms from both samples. Next, a comparison algorithm was applied to these two sets of terms to single out those that were common to both samples, yet whose patterns of occurrences differed between these two samples. Finally, we analyzed and presented this set of terms as content oriented candidates for the predefined topic, in this case "European Politics and Business".

The terms suggested are split into three categories: single word terms, two-word terms and multi-word terms (or phrases). The following three sections describe in detail the methods for generating each of the three categories.

2.4 Suggesting Single Word Terms

Automatically suggesting single word terms as being topically oriented has been most challenging. Our experiments indicated that the "first order" statistics, probability and entropy alone, are not sufficient for gathering information about the topicality of a word in running text. The information in both measurements is essentially equivalent since entropy is just the log inverse of probability.

We found that the "second-order" statistics, such as variance or standard deviation of term frequencies across documents, provide greater insight into topicality. We selected the interval between the occurrences of a word as the basis for analysis. Our intuitions led us to believe that topical single words should appear more frequently and more regularly, i.e. at approximately even intervals, in the focused sample than in the base sample. The focused sample represents, more or less, a topical sublanguage set while the base sample a general language set. Unlike probability and entropy statistics which yield average scores for the whole document, the use of interval makes it possible to get an "instantaneous" measure at any location in the document. More specifically, an interval can be measured "instantaneously" at any point in the text between the occurrences of a particular word. Though using interval alone might still not be sufficient for identifying word topicality, it allowed us to measure the variance which would help identify words that were always changing in their rate of occurrences.

Thus, three scores were generated for each word: the mean log interval, the standard deviation of the mean log interval, and the normalized standard deviation of the mean log interval. The use of a log scale for these measurements is to minimize the effect of unduly large variations in words with long mean intervals. The normalized standard deviation is produced by simply dividing the raw standard deviation by the mean log interval. In most cases, raw standard deviation is found to be larger for words having long mean intervals. In order to compare the standard deviations across words of different intervals, we found this normalization process quite useful.

After scores were generated for all the words in both the focused sample and the base sample, score comparisons between the two samples were carried out in two ways: comparing the intervals and comparing the standard deviations.

To compare the intervals, the "base" mean log interval was subtracted from the "focused" mean log interval and divided by the raw standard deviation from the base sample. The result represents the change of mean log intervals. More explicitly, it yields the number of standard deviations that the "focused" mean log interval is different from the "base" mean log interval. The more negative the value, the more significant the change, and the more prominent the word would appear in the focused sample.

To compare the standard deviations, the normalized "base" standard deviation was subtracted from the normalized "focused" standard deviation. The difference symbolizes how the word is distributed in the focused sample. The more negative the value is, the more "bursty" the word is distributed, and the more likely it is content oriented since "content words tend to appear in 'bursts'" (Church and Mercer 1993).

If a single word term is found in both data samples and it receives negative scores from both interval and standard deviation comparisons, it would be included in the suggested list as being topical oriented.

2.5 Suggesting Two-Word Terms

The method for suggesting two-word terms turned out to be much simpler than that for single word terms though the same techniques are equally applicable. Here, the traditional mutual information score was used. As stated in Church, et al. (1991) and elsewhere, the mutual information measurement can be expressed as:

$$I(w_1 w_2) = \log\left(\frac{p(w_1 w_2)}{p(w_1)\,p(w_2)}\right)$$

where *p(w1w2)* is the frequency in the data collection of the two-word compound (w1,w2); and *p(w1)* and *p(w2)* the frequency of the word constituents. The highest mutual information score indicates that the individual probabilities are low while the two words occur together frequently.

Two steps led to our automatic suggestion of topic-oriented two-word terms. First, the mutual information score was computed for each pair of words that occur in each of the two samples. To capture topicality, we were only interested in pairs of words with high mutual information scores. Therefore, any pair which contained "closed class" words, such as determiners, prepositions, auxiliaries, or single letters, digit numbers, or overly common verbs like "give", "take", etc., were excluded. Such an exclusion not only helped getting pairs of words with high mutual information scores, but also sped up computation significantly. A threshold value was also set such that if any two-word unit occurred less than 3 times in the sample or received a mutual information score lower than 6.0, it was eliminated and would not participate in the next comparison measurement.

With the mutual information scores in hand, a "delta" score was generated by subtracting the "base" mutual information score from the "focused" mutual information score. Topically, prominent two-word terms normally have lower scores in the focused sample that is "keyed" to their topic. This is because the constituent words distribute in wider range of contexts. The probability of them occurring separately increases relative to the probability of them occurring together (Steier and Belew 1994). Therefore, the more negative the "delta" score, the more topically sensitive the two-word term is.

If a two-word term occurs in both data samples and receives a negative "delta" score, it would be included in the suggested list as being topically oriented.

2.6 Suggesting Multi-Word Terms

When automatically suggesting content two-word terms, we looked at the mutual information scores for adjacent words. For multi-word terms, the mutual information score was calculated for non-adjacent words. Our intuitions led us to believe that if there is a significant statistical linkage, i.e. a high mutual information score, between such a pair of words, it is highly possible that they belong to a larger linguistic component.

Our first step was to compute mutual information scores for a word unit separated by a distance of two (i.e. having one unspecified word separating them). Two criteria apply when selecting "interesting" word units. Their mutual information score must be 10 or greater. Following the observations by Steier and Belew (Steier and Belew1994), we only selected pairs which received lower mutual information score in the focused sample than in the base sample.

Once an "interesting" word unit of distance two was selected, a concordance was built of all sentences containing that word unit. These sentences were compared for matching text. If a string of text was found to include that word unit and, at the same time, occur most frequently in the concordance, its leading and trailing "closed-set" words (if any) were chopped off. The remaining text string was presented as a suggested multi-word term.

3. RESULTS and DISCUSSION

3.1 Suggested Single Terms

The focused sample drawn from the 1988 UPI data contains 12,065 unique words. Among them, 5,045 are frequent enough (occurring 3 times or more) to calculate statistics for our experiments (refer to Table 1). The comparison algorithm identified 2,010 suggested terms based on the fact that they received negative scores for both "change of mean log interval" and "distribution burstiness" comparisons. These negative scores indicate that these single word terms have shorter intervals and more regular occurrences in the focused sample.

We compared the suggested list against the single word terms manually selected by the domain expert. The results are summarized in Table 3.

Table 3: Statistics of the Suggested Single Word Terms

	Comparison of Suggested and Manual Terms					
total suggested	total manual	not possible*	no statistics*	possible*	hits	percent included
2,010	276	129	91	56	42	75%

*** not possible: terms not existing in the focused sample**

*** no statistics: terms which have less than 3 occurrences in the focused sample**

*** possible: targeted terms**

Of the 276 topical single terms determined by the domain expert, 129 terms do not exist in the focused sample. As explained earlier, these are the terms intellectually introduced by the domain expert. Almost half of these terms are geographical names in Europe, such as

> albania, albertville, andorra, barcelona, belarus, belorus, bosnia, byelorussia, chancellors, comecon, cp, croatia, erm, eurocurrency, eurofed, europeanization, europeanwide, europeenne, europewide, gaullist, gaullists, gilbraltar, greenland, guemsey, kazakhstan, kirghizia, kirgizia, kyrgystan, kzakhstan, labour, liechtenstein, moldavia, moldova, monaco, nc, nib, nicosia, nuuk, pentagonale, reunify, reykjavik, salzburg, sicily, slovenia, svalbard, tadzhikistan, tajikistan, tajikstan, tirana, tirane, tories, torshavn, turkmenia, turkmenistan, uk, ussr, uzbekistan, vaduz, valletta, weu

Of the remaining 147 actually occurring terms, 91 are not frequent enough to be included in our experiments. They occur in the focused sample two times or less. Again, some of them are geographical names in Europe.

> amsterdam, athens, azerbaijan, bulgaria, estonia, euro, eurodollar, eurodollars, georgia, hamburg, holland, iceland, jersey, latvia, liberals, lithuania, naples, oecd, prague, reunified, rome, russia, serbia, sofia, tory, ukraine, unification

These non-existent and under-represented terms left us with a maximum of 56 terms we could catch in the suggested terms list. Of these, 42 were caught with an accuracy rate of 75% (see Appendix for details).

Further analysis of the missing 14 terms reveals that they were not found in the suggested list due to the statistical constraints we established for our experiments. As shown in Table 4, 13 of these terms received negative scores either for "change of mean log interval" or for "distribution burstiness", but not for both. We believe that their inclusion is possible since they represent what we would call "border-line" suggested terms.

Table 4: "Missed" single word terms

single-word term	dgt1	dgt2	dgt3	dgt4	dgt5	dgt6	dgt7	dgt8
portugal	10	13.75	0.26	16.86	0.23	3.83	-0.81	0.03
europeans	23	12.55	0.35	15.64	0.32	4.98	-0.62	0.03
eec	3	15.49	0.39	19.28	0.32	6.21	-0.61	0.06
luxembourg	12	13.49	0.42	17.06	0.36	6.07	-0.59	0.07
copenhagen	3	15.49	0.47	18.54	0.34	6.23	-0.49	0.14
cyprus	6	14.49	0.44	18.28	0.43	7.89	-0.48	0.01
yugoslavia	12	13.49	0.47	15.33	0.37	5.66	-0.32	0.10
finland	10	13.75	0.51	15.52	0.46	7.19	-0.25	0.05
kgb	5	14.75	0.57	16.41	0.44	7.29	-0.23	0.13
sweden	13	13.38	0.48	14.26	0.44	6.33	-0.14	0.03
turkey	11	13.62	0.53	14.47	0.50	7.25	-0.12	0.03
czechoslovakia	9	13.91	0.09	13.70	0.46	6.29	0.03	-0.36
switzerland	9	13.91	0.21	13.81	0.47	6.48	0.01	-0.26

Statistics Measurements (dgt = digit)

dgt1: number of occurrences (in the focused sample)
dgt2: mean log interval (in the focused sample)
dgt3: normalized SD of mean log interval (in the focused sample)
dgt4: mean log interval (in the base sample)
dgt5: normalized SD of mean log interval (in the base sample)
dgt6: raw SD of mean log interval (in the base sample)
dgt7: ((2nd digit - 4th digit) / 6th digit))
dgt8: (3rd digit - 5th digit)

Admittedly, the suggested list with the total of 2,010 terms is a fairly large one. It obviously contains terms that are not topic oriented. We followed the observations made by Justeson and Katz (1993) and introduced a "post-editing" process. As a result, the list was reduced to 886 terms. Basically, we removed from the original list all the "closed-set" words such as determiners, prepositions, auxiliaries, conjunctions, single letters, etc., as well as other less semantically laden words such as adverbs and verbs.

3.2 Suggested Two-Word Terms

Among 512 "interesting" two-word terms, 170 receive negative "delta" scores. These 164 terms were presented in our suggested two-word terms (see Appendix for details).

A total of 36 topical terms were manually determined based on the UPI focused sample. Of this number, only 26 are actually existent terms, which means that 10 terms were introduced independent of the source material. Among these 26 terms, 6 were too infrequent to generate meaningful statistics though the mutual information scores are high (see Table 5). Five terms, i.e. *E C, U K, the Channel, the Continent,* and *the Wall* failed to participate in statistical screening because they contain "closed-set" words, i.e. single letters and the determiner *the.*

Table 5: "No statistics" two-word terms

two-word term	digit1	digit2
monte carlo	1	13.61674723
social democrats	1	9.58432575
coalition government	1	7.59034954
supreme soviet	1	5.06985277
downing street	1	11.75425075
socialist party	2	6.36709503

Statistical measurements

digit1: frequency (in the focused sample)
digit2: mutual information score

Of the remaining catchable15 two-word terms, 8 are included in the suggested list. Table 6 summarizes the statistics of the suggested two-word terms.

Table 6: Statistics of the Suggested Two-Word Terms

	Comparison of Suggested and Manual Terms					
total suggested	total manual	not possible*	no statistics*	possible*	hits	percent included
170	36	10	11	15	8	53%

*** not possible: terms not existing in the focused sample**
*** no statistics: terms which have less than 3 occurrences in the focused sample**
*** possible: targeted terms**

Further screening revealed that 3 manually selected two-word terms (i.e. *cold war, common market,* and *North Sea*) were actually captured in the 512 "interesting" list. They were not included in the suggested list because they did not receive negative "delta" scores. The suggested list fails to include 4 manually selected two-word terms because their mutual information scores go up. Typically, content oriented two-word terms within the topically related subset of documents are expected to go down. This might be caused by the individual word probabilities. To use Steier and Belew's terms (Steier and Belew 1994), these pairs appear more "opaque", meaning that their constituent words are more probable individually than when they are combined in the focused sample. Table 7 lists these 4 two-word terms appearing in both samples.

Table 7: "Missed" two-word terms

Sample	two-word term	frequency	MI score
“base”	atlantic alliance	11	8.80256520
“focused”	atlantic alliance	4	9.36193333
“base”	cold war	54	8.04486800
“focused”	cold war	11	9.97241419
“base”	common market	26	6.86310460
“focused”	common market	17	7.84030540
“base”	united kingdom	49	7.55353160
“focused”	united kingdom	25	7.80705217

Our suggested two-word terms list (see the Appendix) contains quite a number of useful additional terms about the targeted predefined topic "European Politics and Business". The following are some examples:

US-European relations/politics:

armed forces, diplomatic relations, nuclear missiles, nuclear weapons, trade barriers

European Business:

bilateral trade, economic reform, market integration, private enterprise, private investment

Notable European entities:

banca commerciale, berlin wall, british spies, swiss francs, brussels belgium

Heads of state:

felipe gonzalez, francois mitterrand, mikhail gorbachev

3.3 Suggested Multi-Word Terms

A total of 97 multi-word terms were extracted from the focused sample for inclusion in the suggested list (see Appendix). Admittedly, some of them are simply sentence fragments instead of real phrases.

Of the 35 multi-word terms manually selected by the domain expert, 26 actually occur in the focused sample. As with the single word and two-word terms, the other 9 multi-word terms are simply intellectual introductions from the domain expert. Of the 26 terms, 22 occur frequently enough to generate meaningful statistics. Out of these 22 catchable terms, only 5 are included in the suggested list. Table 8 presents the statistical summary.

Table 8: Statistics of the Suggested Multi-Word Terms

	Comparison of Suggested and Manual Terms					
total suggested	total manual	not possible*	no statistics*	possible*	hits	percent included
97	35	9	4	22	5	23%

*** not possible: terms not existing in the focused sample**

*** no statistics: terms which have less than 3 occurrences in the focused sample**

*** possible: targeted terms**

One possible explanation for not being able to match more manual selections is that most of the two-word terms that could have been used to detect these phrases consist of two common words, such as *house, lords, fund, system.* These two-word terms typically generate fairly low mutual information scores since the constituent words occur frequently by themselves.

It is important to point out that the suggested list does contain a number of useful multi-word terms that are related to the targeted predefined topic "European Politics and Business". For example,

US-European relations/politics:

> short range nuclear missiles, tactical nuclear weapons, conventional arms reduction, multi party system

European Business:

> gross national product, higher interest rates and inflation, Bank of England, North Sea Oil

Notable European entities:

> predominantly Catholic Irish Republic, three British hostages, World War II, Roman Catholic Church

Heads of state or notable dignitaries:

> Secretary of State James Baker, Secretary of State George Shultz, French President Francois Mitterrand, West German Chancellor Helmut Kohl, Soviet leader Mikhail Gorbachev, Soviet Foreign Minister Eduard Shevardnadze

4. CONCLUSION

This paper presents a preliminary experiment in identifying significant terminological units from running text. By comparing a focused sample randomly drawn for a predefined topic against a larger and more general base sample, we can automatically suggest topic-oriented terms based on the detection of significant changes in some statistical measurements. Our experiment on one predefined topic demonstrated that, compared to the manual selection of the topical terms, our suggested lists do contain more useful terms that can be used to describe the topic. We also found that the method is efficient enough for applications to very large textual corpora. Our next step is to further refine the methods by carrying out more experiments across different topics. We mentioned a number of times that our methods were developed based on our intuitive assumptions or hypotheses. More experiments on more topics will prove whether we can obtain positive and consistent results.

Identification of significant terms from running text can be very useful in building intelligent information management systems. Terms identified are good candidates for key word indexing of electronic sources. Topic specificity can assist in grouping or clustering on-line documents. For an information retrieval system, terms identified for a pre-determined subject can be used to develop specialized libraries or files for targeted user groups. Our experiment demonstrated that the methods described can identify various people names, organization entities and other proper names. Those special text tokens are important for constructing text extraction systems.

ACKNOWLEDGEMENTS

This research was done while the second author worked at LEXIS-NEXIS during the summer of 1994. The authors would like to thank Dan Pliske, Mark Wasson and Rob Keefer for helpful comments on this paper, and Rita Freese for proofreading the final draft. The authors were also benefited from numerous conversations with Ken Church at Bell Labs.

REFERENCES

K. Church and P. Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1), March 1990.

K. Church and R. Mercer. Introduction to the special issue in computational linguistics using large corpora. *Computational Linguistics*, 19(1), March 1993.

K. Church, et al. Using statistics in lexical analysis. In U. Zernik, editor, *Lexical Acquisition: Exploring On-line Resources to Build a Lexicon*, Lawrence Erlbaum Association, 1991.

Y. Choueka. Looking for needles in a haystack. *In proceedings, RIAO, Conference on User-Oriented Context Based Text and Image Handling.* Cambridge, MA. 1988.

C. Gierl and D. Frost. Identification of domain-specific terminology by combining mutual information and lexical induction. In B. Neumann, editor, *10th European Conference on Artificial Intelligence.* 1992.

S. Justeson and S. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Research Report. IBM Research Division, T. J. Watson Research Center.* 1993.

F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), March 1993.

A. Steier and R. Belew. Exploring phrases: a statistical analysis of topical language. *Technical Report. University of Califomia - San Diego.* 1994.

D. Walker. Text analysis. In proceedings. *Conference on Applied Natural Language Processing.* 1983

Appedix

Suggested Single Terms which Match the Manual Selections

stalinist	005	14.75	0.10	17.28	0.20	3.40	-0.74	-0.09
european	333	08.70	0.30	11.95	0.39	4.67	-0.70	-0.09
europe	391	08.47	0.31	11.73	0.40	4.69	-0.70	-0.09
netherlands	012	13.49	0.10	15.91	0.23	3.62	-0.67	-0.13
britain	130	10.05	0.33	12.87	0.37	4.81	-0.59	-0.05
unified	013	13.38	0.24	16.31	0.31	5.13	-0.57	-0.07
chancellor	028	12.27	0.21	15.07	0.34	5.17	-0.54	-0.13
budapest	011	13.62	0.31	16.58	0.34	5.64	-0.52	-0.03
belfast	017	12.99	0.40	17.01	0.46	7.84	-0.51	-0.06
nato	140	09.95	0.38	13.14	0.48	6.32	-0.50	-0.10
ireland	035	11.95	0.40	14.98	0.40	6.05	-0.50	-0.01
brussels	020	12.75	0.24	15.14	0.32	4.89	-0.49	-0.09
kremlin	018	12.91	0.31	15.28	0.32	4.93	-0.48	-0.01
ira	062	11.12	0.48	15.15	0.56	8.52	-0.47	-0.08
poland	120	10.17	0.41	12.95	0.46	5.97	-0.47	-0.05
belgium	015	13.17	0.19	15.19	0.30	4.60	-0.44	-0.11
ec	044	11.62	0.46	15.15	0.54	8.11	-0.44	-0.08
denmark	006	14.49	0.23	16.34	0.27	4.36	-0.42	-0.03
hungary	072	10.91	0.40	13.41	0.47	6.24	-0.40	-0.06
vienna	019	12.83	0.20	14.70	0.32	4.76	-0.39	-0.13
greece	011	13.62	0.26	15.76	0.35	5.51	-0.39	-0.09
scotland	010	13.75	0.28	15.67	0.31	4.91	-0.39	-0.04
spain	037	11.87	0.40	14.16	0.42	5.89	-0.39	-0.02
bonn	008	14.08	0.15	16.08	0.34	5.51	-0.36	-0.19
parliament	064	11.08	0.30	12.81	0.38	4.86	-0.36	-0.08
conservatives	012	13.49	0.31	15.22	0.35	5.40	-0.32	-0.04
solidarity	043	11.65	0.39	13.66	0.48	6.56	-0.31	-0.09
reunification	012	13.49	0.39	15.80	0.48	7.51	-0.31	-0.09
malta	019	12.83	0.41	15.15	0.49	7.42	-0.31	-0.08
france	056	11.27	0.27	12.86	0.41	5.24	-0.30	-0.13
germany	112	10.27	0.34	11.89	0.45	5.33	-0.30	-0.11
paris	038	11.83	0.36	13.26	0.37	4.85	-0.30	-0.01
glasnost	007	14.27	0.13	15.89	0.36	5.65	-0.29	-0.22
conservative	028	12.27	0.33	13.56	0.34	4.56	-0.28	-0.01
perestroika	015	13.17	0.37	14.91	0.43	6.34	-0.27	-0.05
dublin	003	15.49	0.24	17.11	0.36	6.14	-0.26	-0.12
helsinki	005	14.75	0.37	16.54	0.42	7.03	-0.25	-0.05
berlin	023	12.55	0.45	14.06	0.46	6.47	-0.23	-0.01
liberal	012	13.49	0.09	14.64	0.36	5.30	-0.22	-0.28
labor	045	11.58	0.32	12.67	0.39	5.00	-0.22	-0.07
organization	037	11.87	0.23	12.50	0.23	2.91	-0.22	-0.01
armenia	004	15.08	0.29	16.82	0.49	8.32	-0.21	-0.21
wales	003	15.49	0.27	16.54	0.37	6.16	-0.17	-0.10
italy	027	12.32	0.27	13.17	0.43	5.64	-0.15	-0.15
austria	015	13.17	0.37	14.26	0.52	7.42	-0.15	-0.15
england	030	12.17	0.30	12.84	0.35	4.52	-0.15	-0.06
imf	005	14.75	0.41	15.76	0.54	8.59	-0.12	-0.13
romania	017	12.99	0.36	13.55	0.48	6.52	-0.09	-0.12
politburo	007	14.27	0.34	14.65	0.45	6.66	-0.06	-0.11
norway	003	15.49	0.04	15.54	0.38	5.89	-0.01	-0.33

Refer to Table 4 for explanations

Suggested Two-Word Terms

```
00005 arms reduction
00005 banca commerciale
00005 barbed wire
00005 british broadcasting
00005 british spies
00005 de mita
00005 developing countries
00005 diplomatic relations
00005 domestic demand
00005 eastern europe
00005 economic activity
00005 french francs
00005 grand jury
00005 hewlett packard
00005 hostile takeover
00005 inflationary pressures
00005 insurance companies
00005 joseph biden
00005 li peng
00005 liberation organization
00005 manufacturers hanover
00005 market integration
00005 monetary fund
00005 nato allies
00005 neil kinnock
00005 oil spill
00005 outlawed ira
00005 palestine liberation
00005 pedro sula
00005 political parties
00005 preferred stock
00005 private enterprise
00005 private investment
00005 sales tax
00005 schering plough
00005 seasonally adjusted
00005 spiritual leader
00005 spokesman marlin
00005 swiss francs
00005 unemployment rate
00005 xinhua news
00006 annual rate
00006 avis europe
00006 bilateral trade
00006 british colony
00006 consumer spending
00006 defense secretary
00006 domestic product
00006 export subsidies
00006 felipe gonzalez
00006 general motors
00006 german marks
00006 gross domestic
00006 latin america
00006 mcdonnell douglas
00006 mti quoted
00006 parent company
00006 price index
00006 real estate
00006 refugee status
00006 soviet union
00006 strategic arms
00006 surged cents
00006 takeover bid
00006 trade representative
00007 british petroleum
00007 british telecom
00007 central bank
00007 democratic party
00007 discount rate
00007 du pont
00007 geoffrey howe
00007 great britain
00007 islamic republic
00007 leveraged buyout
00007 new caledonia
00007 retail sales
00007 ruhollah khomeini
00007 shares changing
00008 aviation administration
00008 ayatollah ruhollah
00008 berlin wall
00008 british columbia
00008 budapest hungary
00008 eduard shevardnadze
00008 foreign affairs
00008 george bush
00008 iron curtain
00008 latin american
00008 lech walesa
00008 marlin fitzwater
00008 minister eduard
00008 roman catholic
00009 exchequer nigel
00009 inf treaty
00009 monetary policy
```

00009	national security
00009	nigel lawson
00009	satanic verses
00009	security forces
00009	tehran radio
00009	tender offer
00009	unbundled units
00009	warsaw poland
00010	armed forces
00010	chemical weapons
00010	grapeseed oil
00010	moderate trading
00010	private sector
00010	queen elizabeth
00010	salman rushdie
00011	del monte
00011	economic reforms
00011	german chancellor
00011	oil prices
00011	pence cents
00011	range missiles
00011	tactical nuclear
00011	trade deficit
00011	van buren
00012	joint ventures
00012	stock prices
00013	chancellor helmut
00013	helmut kohl
00013	middle east
00013	sinn fein
00014	brussels belgium
00014	bush administration
00014	francois mitterrand
00014	rjr nabisco
00015	conventional forces
00015	foreign ministers
00015	joint venture
00015	nuclear missiles
00016	hong kong
00016	trade barriers
00017	british airways
00017	range nuclear
00017	shares compared
00018	world war
00019	conventional arms
00019	foreign policy
00020	stock index
00021	communist party
00024	east germany
00024	gained cents
00025	east bloc
00025	foreign minister
00026	leader mikhail
00026	news conference
00027	news agency
00028	economic growth
00030	interest rates
00030	nuclear weapons
00032	margaret thatcher
00032	warsaw pact
00034	vice president
00034	west german
00036	million shares
00036	minister margaret
00042	mikhail gorbachev
00043	stock exchange
00048	soviet leader
00057	hong kong
00065	west germany
00072	european community
00080	prime minister
00087	soviet union
00099	eastern europe
00912	press international

column 1: number of occurrences (in the focused sample)
column 2: suggested two-word term

Suggested Multi-Word Terms

00001	Conference on Security and Cooperation in Europe
00001	East West relations
00001	Embassy in London
00001	Foreign Relations Committee
00001	Iranian Embassy in London
00001	London Stock Exchange
00001	NATO Secretary General Manfred Woerner
00001	United Arab Emirates
00001	United Press International
00001	United States and Canada
00001	aid to Poland
00001	cents a share
00001	change in Eastern Europe
00001	conference on security

00001	division of Europe
00001	five Central American
00001	foreign direct investment
00001	million shares compared
00001	president and chief
00002	Bank of England
00002	Catholic Irish Republic
00002	Central Statistical Office
00002	Chancellor Helmut Kohl
00002	Chancellor of the Exchequer Nigel Lawson
00002	Civil Aviation Administration
00002	Dow Jones industrial
00002	East West tensions
00002	Embassy in Tehran
00002	Federal Reserve Board
00002	French President Francois Mitterrand
00002	General Electric Co
00002	IRA is fighting to end British rule
00002	Irish Republican Army
00002	New York Stock Exchange
00002	North Sea oil
00002	Palestine Liberation Organization
00002	Poland and Hungary
00002	Roman Catholic Church
00002	San Pedro Sula
00002	Secretary of State George Shultz
00002	Secretary of State James Baker
00002	Soviet made Ilyushin
00002	Stock prices closed
00002	White House spokesman Marlin Fitzwater
00002	World War II
00002	Xinhua news agency
00002	balance of payments
00002	changes in Eastern Europe
00002	chief executive officer
00002	conventional arms reductions
00002	crossed the border
00002	days of talks
00002	gross national product
00002	high interest rates
00002	higher in moderate trading
00002	higher interest rates
00002	higher interest rates and inflation
00002	imposition of martial law
00002	interest rates and inflation
00002	key Financial Times 100 stock index
00002	member of Parliament
00002	missiles with ranges of 300 to 3,400 miles
00002	multi party system
00002	narrower top 30 industrial average gained
00002	offer for Irving
00002	opposition Labor Party
00002	predominantly Catholic Irish Republic
00002	quoted as saying
00002	research and development
00002	secretary of state
00002	senior vice president
00002	shadowy pro Iranian group
00002	short range missiles
00002	short range nuclear
00002	spokesman Marlin Fitzwater
00002	tactical nuclear weapons
00002	tanks and artillery
00002	three British hostages
00002	top 30 industrial average gained
00002	trade and industry
00002	turmoil in China
00003	Ayatollah Ruhollah Khomeini
00004	Dow Jones industrial average
00004	Foreign Minister Eduard Shevardnadze
00004	Islamic Republic News Agency
00004	Prime Minister Felipe Gonzalez
00004	Prime Minister Margaret Thatcher
00004	Soviet Foreign Minister Eduard Shevardnadze
00004	Soviet leader Mikhail Gorbachev
00004	West German Chancellor Helmut Kohl
00004	fighting to end British rule
00004	million shares changing hands
00004	most widely traded stocks
00004	short range nuclear missiles
00005	Civil Aviation Administration of China
00005	North Atlantic Treaty Organization
00006	official Islamic Republic News Agency

column 1: number of occurrences (in the focused sample)

column 2: suggested multi-word term