Automatic Identification of Non-compositional Phrases

Dekang Lin

Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada, R3T 2N2
lindek@cs.umanitoba.ca

and

UMIACS
University of Maryland
College Park, Maryland, 20742
lindek@umiacs.umd.edu

Abstract

Non-compositional expressions present a special challenge to NLP applications. We present a method for automatic identification of non-compositional expressions using their statistical properties in a text corpus. Our method is based on the hypothesis that when a phrase is non-composition, its mutual information differs significantly from the mutual informations of phrases obtained by substituting one of the word in the phrase with a similar word.

1 Introduction

Non-compositional expressions present a special challenge to NLP applications. In machine translation, word-for-word translation of non-compositional expressions can result in very misleading (sometimes laughable) translations. In information retrieval, expansion of words in a non-compositional expression can lead to dramatic decrease in precision without any gain in recall. Less obviously, non-compositional expressions need to be treated differently than other phrases in many statistical or corpus-based NLP methods. For example, an underlying assumption in some word sense disambiguation systems, e.g., (Dagan and Itai, 1994; Li et al., 1995; Lin, 1997), is that if two words occurred in the same context, they are probably similar. Suppose we want to determine the intended meaning of "product" in "hot product". We can find other words that are also modified by "hot" (e.g., "hot car") and then choose the meaning of "product" that is most similar to meanings of these words. However, this method fails when non-compositional expressions are involved. For instance, using the same algorithm to determine the meaning of "line" in "hot line", the words "product", "merchandise", "car", *etc.*, would lead the algorithm to choose the "line of product" sense of "line".

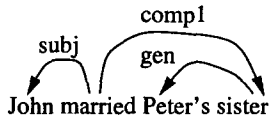
We present a method for automatic identification of non-compositional expressions using their statistical properties in a text corpus. The intuitive idea behind the method is that the metaphorical usage of a non-compositional expression causes it to have a different distributional characteristic than expressions that are similar to its literal meaning.

2 Input Data

The input to our algorithm is a collocation database and a thesaurus. We briefly describe the process of obtaining this input. More details about the construction of the collocation database and the thesaurus can be found in (Lin, 1998).

We parsed a 125-million word newspaper corpus with Minipar,[1] a descendent of Principar (Lin, 1993; Lin, 1994), and extracted dependency relationships from the parsed corpus. A dependency relationship is a triple: `(head type modifier)`, where `head` and `modifier` are words in the input sentence and `type` is the type of the dependency relation. For example, (1a) is an example dependency tree and the set of dependency triples extracted from (1a) are shown in (1b).

(1) a.



b. (marry V:subj:N John), (marry V:comp1:N sister), (sister N:gen:N Peter)

There are about 80 million dependency relationships in the parsed corpus. The frequency counts of dependency relationships are filtered with the log-likelihood ratio (Dunning, 1993). We call a dependency relationship a collocation if its log-likelihood ratio is greater than a threshold (0.5). The number of unique collocations in the resulting database[2] is about 11 million.

Using the similarity measure proposed in (Lin, 1998), we constructed a corpus-based thesaurus[3] consisting of 11839 nouns, 3639 verbs and 5658 adjective/adverbs which occurred in the corpus at least 100 times.

3 Mutual Information of a Collocation

We define the probability space to consist of all possible collocation triples. We use $|H\ R\ M|$ to denote the

[1] available at http://www.cs.umanitoba.ca/~lindek/minipar.htm/
[2] available at http://www.cs.umanitoba.ca/~lindek/nlldemo.htm/
[3] available at http://www.cs.umanitoba.ca/~lindek/nlldemo.htm/

frequency count of all the collocations that match the pattern (H R M), where H and M are either words or the wild card (*) and R is either a dependency type or the wild card. For example,

- |marry V:comp1:N sister| is the frequency count of (marry V:comp1:N sister).
- |marry V:comp1:N *| is the total frequency count of collocations in which the head is marry and the type is V:comp1:N (the verb–object relation).
- |* * *| is the total frequency count of all collocations extracted from the corpus.

To compute the mutual information in a collocation, we treat a collocation (head type modifier) as the conjunction of three events:

A: (* type *)

B: (head * *)

C: (* * modifier)

The mutual information of a collocation is the logarithm of the ratio between the probability of the collocation and the probability of events A, B, and C co-occur if we assume B and C are conditionally independent given A:

(2)

$$\text{mutualInfo(head, type, modifier)} = \log \frac{P(A,B,C)}{P(B|A)P(C|A)P(A)}$$

$$= \log\left(\frac{|\text{head type modifier}| \times |* * *|}{|* \text{ type } *| \times |\text{head type } *| \times |* \text{ type modifier}|}\right)$$

$$= \log\left(\frac{|\text{head type modifier}| \times |* \text{ type } *|}{|\text{head type } *| \times |* \text{ type modifier}|}\right)$$

4 Mutual Information and Similar Collocations

In this section, we use several examples to demonstrate the basic idea behind our algorithm.

Consider the expression "spill gut". Using the automatically constructed thesaurus, we find the following top-10 most similar words to the verb "spill" and the noun "gut":

spill: leak 0.153, pour 0.127, spew 0.125, dump 0.118, pump 0.098, seep 0.096, burn 0.095, explode 0.094, burst 0.092, spray 0.091;

gut: intestine 0.091, instinct 0.089, foresight 0.085, creativity 0.082, heart 0.079, imagination 0.076, stamina 0.074, soul 0.073, liking 0.073, charisma 0.071;

The collocation "spill gut" occurred 13 times in the 125-million-word corpus. The mutual information of this collocation is 6.24. Searching the collocation database, we find that it does not contain any collocation in the form ($simv_{spill}$ V:comp1:N gut) nor (spill V:comp1:N $simn_{gut}$), where $simv_{spill}$ is a verb similar to "spill" and $simn_{gut}$ is a noun similar to "gut". This means that the phrases, such as "leak gut", "pour gut", ... or "spill intestine", "spill instinct", either did not appear in the corpus at all, or did not occur frequent enough to pass the log-likelihood ratio test.

The second example is "red tape". The top-10 most similar words to "red" and "tape" in our thesaurus are:

red: yellow 0.164, purple 0.149, pink 0.146, green 0.136, blue 0.125, white 0.122, color 0.118, orange 0.111, brown 0.101, shade 0.094;

tape: videotape 0.196, cassette 0.177, videocassette 0.168, video 0.151, disk 0.129, recording 0.117, disc 0.113, footage 0.111, recorder 0.106, audio 0.106;

The following table shows the frequency and mutual information of "red tape" and word combinations in which one of "red" or "tape" is substituted by a similar word:

Table 1: red tape

verb	object	freq	mutual info
red	tape	259	5.87
yellow	tape	12	3.75
orange	tape	2	2.64
black	tape	9	1.07

Even though many other similar combinations exist in the collocation database, they have very different frequency counts and mutual information values than "red tape".

Finally, consider a compositional phrase: "economic impact". The top-10 most similar words are:

economic: financial 0.305, political 0.243, social 0.219, fiscal 0.209, cultural 0.202, budgetary 0.2, technological 0.196, organizational 0.19, ecological 0.189, monetary 0.189;

impact: effect 0.227, implication 0.163, consequence 0.156, significance 0.146, repercussion 0.141, fallout 0.141, potential 0.137, ramification 0.129, risk 0.126, influence 0.125;

The frequency counts and mutual information values of "economic impact" and phrases obtained by replacing one of "economic" and "impact" with a similar word are in Table 4. Not only many combinations are found in the corpus, many of them have very similar mutual information values to that of

Table 2: economic impact

verb	object	freq	mutual info
economic	impact	171	1.85
financial	impact	127	1.72
political	impact	46	0.50
social	impact	15	0.94
budgetary	impact	8	3.20
ecological	impact	4	2.59
economic	effect	84	0.70
economic	implication	17	0.80
economic	consequence	59	1.88
economic	significance	10	0.84
economic	fallout	7	1.66
economic	repercussion	7	1.84
economic	potential	27	1.24
economic	ramification	8	2.19
economic	risk	17	-0.33

"economic impact". In fact, the difference of mutual information values appear to be more important to the phrasal similarity than the similarity of individual words. For example, the phrases "economic fallout" and "economic repercussion" are intuitively more similar to "economic impact" than "economic implication" or "economic significance", even though "implication" and "significance" have higher similarity values to "impact" than "fallout" and "repercussion" do.

These examples suggest that one possible way to separate compositional phrases and non-compositional ones is to check the existence and mutual information values of phrases obtained by substituting one of the words with a similar word. A phrase is probably non-compositional if such substitutions are not found in the collocation database or their mutual information values are significantly different from that of the phrase.

5 Algorithm

In order to implement the idea of separating non-compositional phrases from compositional ones with mutual information, we must use a criterion to determine whether or not the mutual information values of two collocations are significantly different. Although one could simply use a predetermined threshold for this purpose, the threshold value will be totally arbitrary. Furthermore, such a threshold does not take into account the fact that with different frequency counts, we have different levels confidence in the mutual information values.

We propose a more principled approach. The frequency count of a collocation is a random variable with binomial distribution. When the frequency count is reasonably large (e.g., greater than 5), a binomial distribution can be accurately approximated by a normal distribution (Dunning, 1993). Since all the potential non-compositional expressions that we are considering have reasonably large frequency counts, we assume their distributions are normal.

Let $|\texttt{head type modifier}| = k$ and $|* * *| = n$. The maximum likelihood estimation of the true probability p of the collocation (`head type modifier`) is $\tilde{p} = \frac{k}{n}$. Even though we do not know what p is, since p is (assumed to be) normally distributed, there is N% chance that it falls within the interval

$$\frac{k}{n} \pm z_N \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} = \frac{k}{n} \pm z_N \frac{\sqrt{k(1-\frac{k}{n})}}{n} \approx \frac{k \pm z_N \sqrt{k}}{n}$$

where z_N is a constant related to the confidence level N and the last step in the above derivation is due to the fact that $\frac{k}{n}$ is very small. Table 3 shows the z_N values for a sample set of confidence intervals.

Table 3: Sample z_N values

N%	50%	80%	90%	95%	98%	99%
z_N	0.67	1.28	1.64	1.96	2.33	2.58

We further assume that the estimations of $P(A)$, $P(B|A)$ and $P(C|A)$ in (2) are accurate. The confidence interval for the true probability gives rise to a confidence interval for the true mutual information (mutual information computed using the true probabilities instead of estimations). The upper and lower bounds of this interval are obtained by substituting $\frac{k}{n}$ with $\frac{k+z_N\sqrt{k}}{n}$ and $\frac{k-z_N\sqrt{k}}{n}$ in (2). Since our confidence of p falling between $\frac{k \pm z_N \sqrt{k}}{n}$ is N%, we can have N% confidence that the true mutual information is within the upper and lower bound.

We use the following condition to determine whether or not a collocation is compositional:

(3) A collocation α is non-compositional if there does not exist another collocation β such that (a) β is obtained by substituting the head or the modifier in α with a similar word and (b) there is an overlap between the 95% confidence interval of the mutual information values of α and β.

For example, the following table shows the frequency count, mutual information (computed with the most likelihood estimation) and the lower and upper bounds of the 95% confidence interval of the true mutual information:

verb-object	freq. count	mutual info	lower bound	upper bound
make difference	1489	2.928	2.876	2.978
make change	1779	2.194	2.146	2.239

Since the intervals are disjoint, the two collocations are considered to have significantly different mutual information values.

6 Evaluation

There is not yet a well-established methodology for evaluating automatically acquired lexical knowledge. One possibility is to compare the automatically identified relationships with relationships listed in a manually compiled dictionary. For example, (Lin, 1998) compared automatically created thesaurus with the WordNet (Miller et al., 1990) and Roget's Thesaurus. However, since the lexicon used in our parser is based on the WordNet, the phrasal words in WordNet are treated as a single word. For example, "take advantage of" is treated as a transitive verb by the parser. As a result, the extracted non-compositional phrases do not usually overlap with phrasal entries in the WordNet. Therefore, we conducted the evaluation by manually examining sample results. This method was also used to evaluate automatically identified hyponyms (Hearst, 1998), word similarity (Richardson, 1997), and translations of collocations (Smadja et al., 1996).

Our evaluation sample consists of 5 most frequent open class words in the our parsed corpus: {have, company, make, do, take} and 5 words whose frequencies are ranked from 2000 to 2004: {path, lock, resort, column, gulf}. We examined three types of dependency relationships: object-verb, noun-noun, and adjective-noun. A total of 216 collocations were extracted, shown in Appendix A.

We compared the collocations in Appendix A with the entries for the above 10 words in the NTC's English Idioms Dictionary (henceforth NTC-EID) (Spears and Kirkpatrick, 1993), which contains approximately 6000 definitions of idioms. For our evaluation purposes, we selected the idioms in NTC-EID that satisfy both of the following two conditions:

(4) a. the head word of the idiom is one of the above 10 words.

b. there is a verb-object, noun-noun, or adjective-noun relationship in the idiom and the modifier in the phrase is not a variable. For example, "take a stab at something" is included in the evaluation, whereas "take something at face value" is not.

There are 249 such idioms in NTC-EID, 34 of which are also found in Appendix A (they are marked with the '+' sign in Appendix A). If we treat the 249 entries in NTC-EID as the gold standard, the precision and recall of the phrases in Appendix A are shown in Table 4. To compare the performance with manually compiled dictionaries, we also compute the precision and recall of the entries in the Longman Dictionary of English Idioms (LDOEI) (Long and Summers, 1979) that satisfy the two conditions in (4). It can be seen that the overlap between manually compiled dictionaries are quite low, reflecting the fact that different lexicographers may have quite different opinion about which phrases are non-compositional.

	Precision	Recall	Parser Errors
Appendix A	15.7%	13.7%	9.7%
LDOEI	39.4%	20.9%	N.A.

Table 4: Evaluation Results

The collocations in Appendix A are classified into three categories. The ones marked with '+' sign are found in NTC-EID. The ones marked with '×' are parsing errors (we retrieved from the parsed corpus all the sentences that contain the collocations in Appendix A and determine which collocations are parser errors). The unmarked collocations satisfy the condition (3) but are not found in NTC-EID. Many of the unmarked collocation are clearly idioms, such as "take (the) Fifth Amendment" and "take (its) toll", suggesting that even the most comprehensive dictionaries may have many gaps in their coverage. The method proposed in this paper can be used to improve the coverage manually created lexical resources.

Most of the parser errors are due to the incompleteness of the lexicon used by the parser. For example, "opt" is not listed in the lexicon as a verb. The lexical analyzer guessed it as a noun, causing the erroneous collocation "(to) do opt". The collocation "trig lock" should be "trigger lock". The lexical analyzer in the parser analyzed "trigger" as the -er form of the adjective "trig" (meaning well-groomed).

Duplications in the corpus can amplify the effect of a single mistake. For example, the following disclaimer occurred 212 times in the corpus.

> "Annualized average rate of return after expenses for the past 30 days: not a forecast of future returns"

The parser analyzed "a forecast of future returns" as [S [NP a forecast of future] [VP returns]]. As a result, `(return V:subj:N forecast)` satisfied the condition (3).

Duplications can also skew the mutual information of correct dependency relationships. For example, the verb–object relationship between "take" and "bride" passed the mutual information filter because there are 4 copies of the article containing this phrase. If we were able to throw away the duplicates and record only one count of "take–bride", it would have not pass the mutual information filter (3).

The fact that systematic parser errors tend to pass the mutual information filter is both a curse and a blessing. On the negative side, there is no obvious way to separate the parser errors from true non-compositional expressions. On the positive side, the output of the mutual information filter has much higher concentration of parser errors than the database that contains millions of collocations. By manually sifting through the output, one can construct a list of frequent parser errors, which can then be incorporated into the parser so that it can avoid making these mistakes in the future. Manually going through the output is not unreasonable, because each non-compositional expression has to be individually dealt with in a lexicon anyway.

To find out the benefit of using the dependency relationships identified by a parser instead of simple co-occurrence relationships between words, we also created a database of the co-occurrence relationship between part-of-speech tagged words. We aggregated all word pairs that occurred within a 4-word window of each other. The same algorithm and similarity measure for the dependency database are used to construct a thesaurus using the co-occurrence database. Appendix B shows all the word pairs that satisfies the condition (3) and that involve one of the 10 words {have, company, make, do, take, path, lock, resort, column, gulf}. It is clear that Appendix B contains far fewer true non-compositional phrases than Appendix A.

7 Related Work

There have been numerous previous research on extracting collocations from corpus, e.g., (Choueka, 1988) and (Smadja, 1993). They do not, however, make a distinction between compositional and non-compositional collocations. Mutual information has often been used to separate systematic associations from accidental ones. It was also used to compute the distributional similarity between words (Hindle, 1990; Lin, 1998). A method to determine the compositionality of verb-object pairs is proposed in (Tapanainen et al., 1998). The basic idea in there is that "if an object appears only with one verb (of few verbs) in a large corpus we expect that it has an idiomatic nature" (Tapanainen et al., 1998, p.1290). For each object noun o, (Tapanainen et al., 1998) computes the distributed frequency $DF(o)$ and rank the non-compositionality of o according to this value. Using the notation introduced in Section 3, $DF(o)$ is computed as follows:

$$DF(o) = \sum_{i=1}^{n} \frac{|v_i,\ \texttt{V:comp1:N},\ o|^a}{n^b}$$

where $\{v_1, v_2, \ldots, v_n\}$ are verbs in the corpus that took o as the object and where a and b are constants.

The first column in Table 5 lists the top 40 verb-object pairs in (Tapanainen et al., 1998). The "mi" column show the result of our mutual information filter. The '+' sign means that the verb-object pair is also consider to be non-compositional according to mutual information filter (3). The '–' sign means that the verb-object pair is present in our dependency database, but it does not satisfy condition (3). For each '–' marked pairs, the "similar collocation" column provides a similar collocation with a similar mutual information value (i.e., the reason why the pair is not consider to be non-compositional). The '$\diamond$' marked pairs are not found in our collocation database for various reasons. For example, "finish seventh" is not found because "seventh" is normalized as "_NUM", "have a go" is not found because "a go" is not an entry in our lexicon, and "take advantage" is not found because "take advantage of" is treated as a single lexical item by our parser. The $\surd$ marks in the "ntc" column in Table 5 indicate that the corresponding verb-object pairs is an idiom in (Spears and Kirkpatrick, 1993). It can be seen that none of the verb-object pairs in Table 5 that are filtered out by condition (3) is listed as an idiom in NTC-EID.

8 Conclusion

We have presented a method to identify non-compositional phrases. The method is based on the assumption that non-compositional phrases have a significantly different mutual information value than the phrases that are similar to their literal meanings. Our experiment shows that this hypothesis is generally true. However, many collocations resulted from systematic parser errors also tend to posses this property.

Acknowledgements

The author wishes to thank ACL reviewers for their helpful comments and suggestions. This research was partly supported by Natural Sciences and Engineering Research Council of Canada grant OGP121338.

References

Y. Choueka. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO Conference on User-Oriented Content-Based Text and Image Handling*, Cambridge, MA, March 21-24.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, March.

Marti A. Hearst. 1998. Automated discovery of wordnet relations. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 131–151. MIT Press.

Table 5: Comparison with (Tapanainen et al., 1998)

verb-object	mi	ntc	similar collocation
take toll	+		
go bust	+		
make plain	+		
mark anniversary	−		celebrate anniversary
finish seventh	◇		
make inroad	−		make headway
do homework	−		do typing
have hesitation	−		have misgiving
give birth	+	√	
have a=go	◇	√	
make mistake	−		make miscalculation
go so=far=as	◇		
take precaution	+		
look as=though	◇		
commit suicide	−		commit crime
pay tribute	−		pay homage
take place	+	√	
make mockery	+		
make headway	−		make inroad
take wicket	◇		
cost $	−		cost million
have qualm	−		have misgiving
make pilgrimage	−		make foray
take advantage	◇	√	
make debut	+		
have second=thought	◇	√	
do job	−		do work
finish sixth	◇		
suffer heartattack	◇		
decide whether	◇		
have impact	−		have effect
have chance	−		have opportunity
give warn	◇		
have sexual=intercourse	−		have sex
take plunge	+		
have misfortune	−		share misfortune
thank goodness	+		
have nothing	◇		
make money	−		make profit
strike chord	+	√	

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of ACL-90*, pages 268–275, Pittsburg, Pennsylvania, June.

Xiaobin Li, Stan Szpakowicz, and Stan Matwin. 1995. A WordNet-based algorithm for word sense disambiguation. In *Proceedings of IJCAI-95*, pages 1368–1374, Montreal, Canada, August.

Dekang Lin. 1993. Principle-based parsing without overgeneration. In *Proceedings of ACL-93*, pages 112–120, Columbus, Ohio.

Dekang Lin. 1994. Principar—an efficient, broad-coverage, principle-based parser. In *Proceedings of COLING-94*, pages 482–488. Kyoto, Japan.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACL/EACL-97*, pages 64–71, Madrid, Spain, July.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING/ACL-98*, pages 768–774, Montreal.

T. H. Long and D. Summers, editors. 1979. *Longman Dictionary of English Idioms*. Longman Group Ltd.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Stephen D. Richardson. 1997. *Determining Similarity and Inferring Relations in a Lexical Knowledge Base*. Ph.D. thesis, The City University of New York.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collcations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, March.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–178.

R. A. Spears and B. Kirkpatrick. 1993. *NTC's English Idioms Dictionary*. National Textbook Company.

Pasi Tapanainen, Jussi Piitulainen, and Timo Jävinen. 1998. Idiomatic object usage and support verbs. In *Proceedings of COLING/ACL-98*, pages 1289–1293, Montreal, Canada.

Appendix A

Among the collocations in which the head word is one of {have, company, make, do, take, path, lock, resort, column, gulf}, the 216 collocations in the following table are considered by our program to be idioms (i.e., they satisfy condition (3)). The codes in the remark column are explained as follows:

×: parser errors;

+: collocations found in NTC-EID.

collocation	remark
(to) have (the) decency	
(to) have (all the) earmark(s)	
(to) have enough	+
(to) have falling	+
have figuring	×
have giving	×
(to) have (a) lien (against)	
(to) have (all the) making(s) (of)	
(to) have plenty	
(to) have (a) record	
have working	×
have wrought	×
(a) holding company	
(a) touring company	
(a) insurance company	
Sinhalese make	×
mistake make	×
mos make	×
(to) make abrasive	
(to) make acquaintance	
(to) make believer (out of)	
(to) make bow	
(to) make (a) case	
(to) make (a) catch	
(to) make (a) dash	
(to) make (one's) debut	
(to) make (up) (the) Dow Jones Industrial Average	
(to) make (a) duplicate	
(to) make enemy	
(to) make (an) error	
(to) make (an) exception	+
(to) make (an) excuse	
(to) make (a) fool	+
(to) make (a) fortune	
(to) make friend	+

collocation	remark
(to) make (a) fuss	+
(to) make (a) grab	
(to) make grade	+
(to) make (a) guess	
(to) make hay	+
(to) make headline(s)	
(to) make (a) killing	+
(to) make (a) living	+
(to) make (a) long-distance call	
(to) make (one's) mark	
(to) make (no) mention	
(to) make (one's) mind (up)	+
(to) make (a) mint	
(to) make (a) mockery (of)	
(to) make noise	
(to) make (a) pitch	+
(to) make plain	×
(to) make (a) point	+
(to) make preparation(s)	
(to) make (no) pretense	
(to) make (a) pun	
(to) make referral(s)	
(to) make (the) round(s)	
(to) make (a) run (at)	+
(to) make savings and loan association	×
(to) make (no) secret	
(to) make (up) sect	
(to) make sense	+
(to) make (a) shamble(s) (of)	
(to) make (a) showing	
(to) make (a) splash	
(to) make (a) start	
(to) make (a) stop	
(to) make (a) tackle	
(to) make (a) turn	
(to) make (a) virtue (of)	
(to) make wonder	×
(to) do (an) about-face	+
(to) do at home	×
(to) do bargain-hunting	
(to) do both	
(to) do business	
(to) do (a) cameo	
(to) do casting	
(to) do damage	
(to) do deal(s)	
(to) do (the) deed	
(to) do (a) disservice	
(to) do either	
(to) do enough	
(to) do (a) favor	
(to) do honor(s)	+
(to) do I.	×
(to) do (an) imitation	
(to) do justice	+
(to) do OK	
(to) do opt	×
(to) do puzzle	
do Santos	×
(to) do stunt(s)	
(to) do (the) talking	

collocation	remark
(to) do (the) trick	+
(to) do (one's) utmost (to)	
(to) do well	
(to) do wonder(s)	
(to) do (much) worse	
do you	×
(the) box-office take	
(to) take aim	+
(to) take back	×
(to) take (the) bait	
(to) take (a) beating	
(to) take (a) bet	
(to) take (a) bite	
(to) take (a) bow	+
(to) take (someone's) breath (away)	+
(to) take (the) bride (on honeymoon)	
(to) take charge	+
(to) take command	
(to) take communion	
(to) take countermeasure	
(to) take cover	
(to) take (one's) cue	
(to) take custody	
(to) take (a) dip	
(to) take (a) dive	
(to) take (some) doing	
(to) take (a) drag	
(to) take exception	+
(to) take (the Gish Road) exit	
(to) take (the) factor (into account)	
(to) take (the) Fifth Amendment	
(to) take forever	
(to) take (the) form (of)	
(to) take forward	×
(to) take (a) gamble	
(to) take (a) genius (to figure out)	
(to) take (a) guess	
(to) take (the) helm	
(to) take (a) hit	
(to) take (a) holiday	
(to) take (a) jog	
(to) take knock(s)	
(to) take a lap	
(to) take (the) lead	
(to) take (the) longest	
(to) take (a) look	+
(to) take lying	×
(to) take measure	
(to) take (a) nosedive	+
(to) take note (of)	+
(to) take oath	
(to) take occupancy	
(to) take part	+
(to) take (a) pick	
(to) take place	+
(to) take (a) pledge	
(to) take plunge	
(to) take (a) poke (at)	
(to) take possession	
(to) take (a) pounding	
(to) take (the) precaution(s)	

collocation	remark
(to) take private	×
(to) take profit	
(to) take pulse	
(to) take (a) quiz	
(to) take refuge	
(to) take root	+
(to) take sanctuary	
(to) take seconds	
(to) take shape	
(to) take (a) shine	+
(to) take side(s)	+
(to) take (a) sip	
(to) take (a) snap	
(to) take (the) sting (out of)	
(to) take (12) stitch(es)	
(to) take (a) swing (at)	
(to) take (its) toll	
(to) take (a) tumble	
(to) take (a) turn	+
(to) take (a) vote	
(to) take (a) vow	
(to) take whatever	
(a) beaten path	
mean path	×
(a) career path	
(a) flight path	
(a) garden path	
(a) growth path	
(an) air lock	
(a) power lock	
(a) trig lock	×
(a) virtual lock	
(a) combination lock	
(a) door lock	
(a) rate lock	
(a) safety lock	
(a) shift lock	
(a) ship lock	
(a) window lock	
(to) lock horns	+
(to) lock key	
(a) last resort	
(a) christian resort	
(a) destination resort	
(an) entertainment resort	
(a) ski resort	
(a) spinal column	
(a) syndicated column	
(a) change column	
(a) gossip column	
(a) Greek column	
(a) humor column	
(the) net-income column	
(the) society column	
(the) steering column	
(the) support column	
(a) tank column	
(a) win column	
(a) stormy gulf	

collocation by proximity
have[V] impact[N]
have[V] legend[N]
have[V] Magellan[N]
have[V] midyear[N]
have[V] orchestra[N]
have[V] precinct[N]
have[V] quarter[N]
have[V] shame[N]
have[V] year end[N]
have[V] zoo[N]
mix[N] company[N]
softball[N] company[N]
electronic[A] make[N]
lost[A] make[N]
no more than[A] make[N]
sure[A] make[N]
circus[N] make[N]
flaw[N] make[N]
recommendation[N] make[N]
shortfall[N] make[N]
way[N] make[N]
make[V] arrest[N]
make[V] mention[N]
make[V] progress[N]
make[V] switch[N]
do[V] Angolan[N]
do[V] damage[N]
do[V] FSX[N]
do[V] hair[N]
do[V] harm[N]
do[V] interior[N]
do[V] justice[N]
do[V] prawn[N]
do[V] worst[N]
place[N] take[N]
take[V] precaution[N]
moral[A] path[N]
temporarily[A] path[N]
Amtrak[N] path[N]
door[N] path[N]
reconciliation[N] path[N]
trolley[N] path[N]
up[A] lock[N]
barrel[N] lock[N]
key[N] lock[N]
love[N] lock[N]
step[N] lock[N]
lock[V] Eastern[N]
lock[V] nun[N]
complex[A] resort[N]
international[N] resort[N]
Taba[N] resort[N]
desk-top[A] column[N]
incorrectly[A] column[N]
income[N] column[N]
smoke[N] column[N]
resource[N] gulf[N]
stream[N] gulf[N]

Appendix B (results obtained without a parser)

collocation by proximity
have[V] B[N]
have[V] companion[N]
have[V] conversation[N]
have[V] each[N]