Extracting Verb-Noun Collocations from Text

Jia Yan Jian
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, Taiwan
g914339@oz.nthu.edu.tw

Abstract

In this paper, we describe a new method for extracting monolingual collocations. The method is based on statistical methods extracts VN collocations from large textual corpora. Being able to extract a large number of collocations is very critical to machine translation and many other application. The method has an element of snowballing in it. Initially, one identifies a pattern that will produce a large portion of VN collocations. We experimented with an implementation of the proposed method on a large corpus with satisfactory results. The patterns are further refined to improve on the precision ration.

1 Introduction

Collocations are recurrent combinations of words that co-occur more often than chance. Collocations like terminology tend to be lexicalized and have a somehow more restricted meaning than the surface form suggested (Justerson and Katz 1994). The words in a collocation may be appearing next to each other (rigid collocation) or otherwise (flexible/elastic collocations). On the other hand, collocations can be classified into lexical and grammatical collocations (Benson, Benson, Ilson, 1986). Lexical collocations are formed between content words, while the grammatical collocation has to do with a content word with a function word or a syntactic structure. Collocations are pervasive in all types of writing and can be found in phrases, chunks, proper names, idioms, and terminology.

Automatic extraction of monolingual and bilingual collocations are important for many applications, including Computer Assisted Language Learning, natural language generation, word sense disambiguation, machine translation, lexicography, and cross language information retrieval. Hank and Church (1990) pointed out the usefulness of pointwise mutual information for identifying collocations in lexicography. Justeson and Katz (1995) proposed to identify technical terminology based on preferred linguistic patterns and discourse property of repetition. Among many general methods presented in Manning and Schutze (1999), the best method is filtering based on both linguistic and statistical constraints. Smadja (1993) presented a program called XTRACT, based on mean and variance of the distance between two words that is capable of computing flexible collocations. Kupiec (1992) proposed to extract bilingual noun phrases using statistical analysis of cooccurrence of phrases. Smadja, McKeown, and Hatzivassiloglou (1996) extended the EXTRACT approach to handling of bilingual collocation based mainly on the statistical measures of Dice coefficient. Dunning (1993) pointed out the weakness of mutual information and showed that log likelihood ratios are more effective in identifying monolingual collocations especially when the occurrence count is very low.

Smadja’s XTRACT is the seminal work on extracting collocation types. XTRACT involves three different statistical measures related to how likely a pair of words is part of a collocation type. It is complicated to set different thresholds for each of these statistical measures. We decided to research and develop a new and simpler method for extracting monolingual collocations. We describe the experiments and evaluation in Section 3. The limitations and related issues will be taken up in Section 4. We conclude and give future direction in Section 5.
2 The algorithm

We used Sinorama Corpus to develop methods for extracting monolingual collocations. A number of necessary preprocessing steps were carried out. Those preprocessing steps include:
1. Part of speech tagging for English and Chinese test
2. N-gram construction
3. Logarithmic likelihood ratio (LLR) computation

2.1 Extraction of English VN collocations

In our research, we discovered some problems about XTRACT. The problems with XTRACT include:
1. XTRACT produce a list of collocation types rather than instances.
2. XTRACT is complicated because it requires thresholds for three statistical measures.
3. There is no systematic way of setting thresholds for a certain level of confidence.
4. XTRACT is based on the author’s intuition about collocation.
5. XTRACT does not provide explicitly types of collocation.

For the above reasons, we decided to research and explore new methods for extracting monolingual collocations.

2.1.1 Step1: Computing such VN types with high counts

The method has an element of snowballing in it. Initially, one identifies a pattern that will produce a large portion of VN collocation. We started with the following pattern (1):

\[ \text{V + ART or POSS + \cdots + N} \]  

By extracting such VN types with high counts, we got a list of highly likely collocation types. In addition, we also take the passive form (2) of VN into consideration:

\[ \text{ART or POSS + N + \cdots + be + Ved (the passive VN)} \]  

The list is further filtered for higher precision: the pairs with LLR lower than 7.88 (confidence level 95%) are removed from consideration.

2.1.2 Step2: Extracting VN patterns from corpus

After obtaining the list, we gather all the instances where the VN appears in the corpus. From the instances, we compute the following patterns (3) for extracting VN collocations:

\begin{align*}
\text{POS preceding V} \\
\text{POS sequence between V and O} \\
\text{POS following O}
\end{align*}

and we also consequently consider the passive form and its context:

\[
\text{Log-likelihood ratio : } \text{LLR}(x;y) = -2 \log \frac{p^x (1-p)^{y-x}}{p^x (1-p)^{y-x}}
\]

\[
k_1 : \# \text{ of pairs that contain x and y simultaneously.}
\]

\[
k_2 : \# \text{ of pairs that contain x but do not contain y.}
\]

\[
n_1 : \# \text{ of pairs that contain y}
\]

\[
n_2 : \# \text{ of pairs that does not contain y}
\]

\[
p_1 = k_1/n_1, \quad p_2 = k_2/n_2, \quad p = (k_1+k_2)/(n_1+n_2)
\]
2.1.3 Step3: Manipulating the correct structure statistics of VN patterns

We eliminated patterns that appear less than three times. These patterns are much more stringent than pattern we started out with. These patterns help us get rid of unlikely VN instances such as “make film” in “make a leap into TV and film,” since the POS sequence of “a leap into TV and” has a low count in the initial batch of “likely” collocations. On the other hand, “make film” in “make my first film” would be kept as a legitimate instance of VN, since the pos sequence of “my first” has rather high count in the initial batch of “likely” collocations.

Actually, the POS sequences of intervening words has a skew distribution concentrating on a dozen of short phrases (see Table1):

| VN collocation | Translation   | POS of VN |
|----------------|---------------|-----------|
| ride a bike    | 像自行車      | vb + at + nn |
| take my advice | 聽我的勸告    | vb + pp$ + nn |
| keep a diary   | 習近平      | vb + at + nn |
| action will be taken | 採取行動 | nn + md + be + vbd |
| problem is solved | 解決問題 | nn + be + vbd |
| decision can be made | 做決定 | nn + md + be + vbd |

These patterns can be coupled with other constraints for best results:

1. No punctuation marks should come between V and O
2. The noun closest to the verb takes precedence

For now, we only consider verbs with two obligatory arguments of subject and object. Therefore, we exclude instance like (make, choice) in “make entertainment at home a choice.” We plan to extract VN in three-argument proposition separately.

The other issue has to do with data sparseness. For collocation types with low count, the estimation of LLR is not as reliable. In the future, we will also experiment with using search engine such as Google to estimate word counts and VN instance count for more reliable estimation of LLR.

XTRACT does not touch on the issue of identify VN collocation instances in (6) and exclude that in (5). In our research, we explored the identification of collocation instances and attempt to avoid cases that maybe a correct collocation type but not a correct collocation instance.

… make a leap into TV and film…  (5)
… made great efforts to promote documentary film…  (6)

2.2 Example

To extract VN collocations, we first run part of speech tagging on sentences. For instance, we get the results of tagging below:

He/pps defines/vbz success/nn for/in a/at paper/nn as/cs not/* needing/vbg to/to exert/vb political/jj influence/nn or/cc obtain/vb financial/jj subsidies/nns ,/, but/cc rather/rb being/beg able/jj to/to rely/vb wholly/rb on/in content/nn to/to attract/vb readers/nns that/cs in/in turn/nn attract/vb advertisers/nns ,/, and/cc thus/rb keep/vb afloat/rb by/in its/pp$ own/jj efforts/nns ./.
After tagging English sentences, we construct N-gram extracted likely VN types with high count from bigram, trigram and fourgram. We then obtained got a list of highly likely collocation types (Table 2). The pairs with LLR lower then 7.88 are eliminated from Table 2. If the pair appeared less than once. we also eliminated the pair.

After obtaining likely collocation types, we gathered all instances where the VN appears in the corpus. The distance between the verb and the object is at most five words. Both of the words before the verb and after the object are recorded. Table 3 shows those patterns of VN instances.

### Table 2
A list of highly likely collocation types

| Verb   | Noun   | Count (VN) | Count(V) | Count(N) | llr_score |
|--------|--------|------------|----------|----------|-----------|
| have   | influence | 24         | 5293     | 57       | 52.28961  |
| exert  | influence | 4          | 14       | 57       | 40.58210  |
| exercise | influence | 4          | 23       | 57       | 36.09338  |
| reduce | influence | 3          | 188      | 57       | 12.43681  |
| eradicate | influence | 1         | 6        | 57       | 8.876641  |
| root   | influence | 1          | 6        | 57       | 8.876641  |

### Table 3
Extracting VN collocation from corpus

| Rec | V-1 | Verb | N-5 | N-4 | N-3 | N-2 | N-1 | Noun | N+1 |
|-----|-----|------|-----|-----|-----|-----|-----|------|-----|
| 96335 | t' | have |     |     |     |     | much | influence | on |
| 55203 | woman | have |     |     |     |     | some | influence | . |
| 129530 | tank | have |     |     |     | a | considerable | influence | on |
| 122706 | He | have |     |     |     | an | influence | on |
| 123975 | mother | have |     |     |     |     | considerable | influence | . |
| 125192 | Wen | have |     |     |     | a | great | influence | on |
| 9326 | which | have |     |     |     | such | a | powerful | influence | on |
| 56033 | as | have |     |     |     | an | enormous | influence | throughout |
| 67666 | have | have |     |     |     | less | influence | than |
| 76130 | have | have |     |     |     | lasting | influence | on |
| 95098 | always | have |     |     |     | a | certain | influence | on |
| 125182 | Xi | have |     |     |     | the | greatest | influence | on |
| 5704 | have | have |     |     |     | a | very | negative | influence | . |
| 1742 | have | have |     |     | a | deep and | lasting | influence | . |
| 111368 | owner | have |     |     |     | no | less | influence | than |
| 96654 | thus | have |     |     |     | a | decisive | influence | on |
| 109816 | family | have |     |     |     | the | greatest | influence | on |
| 115428 | png | have |     |     | be | under | foreign | influence | . |
| 39165 | to | exert |     |     |     |     |     | influence | . |
| 112540 | to | exert |     |     |     |     | political | influence | or |
| 118754 | to | exert |     |     |     |     | his | influence | to |
| 106807 | to | exert |     |     |     | a | positive | influence | for |
| 106846 | it | exert | a | powerful | cultural | influence | throughout |
| 46061 | whohas | exert |     |     |     |     | enormous | influence | upon |
| 123962 | best | exercise |     |     |     | a | restrain | influence | on |
| 40774 | and | exercise |     |     |     | her | political | influence | in |
| 127061 | to | reduce |     |     |     |     | the | influence | of |
3 Experiment and evaluation

We worked with around 50,000 aligned sentences from the Sinorama parallel Corpus in our experiments with an implementation of the proposed method. The average English sentence had 43.95 words. From the experimental data, we have extracted 17,298 VN collocation types. Then, we could obtain 45,080 VN instances for these VN types. See Table 3 for some examples for the verb “influence.”

We select 100 sentences from the parallel corpus of Sinorama magazine to evaluate the performance. A human judge majoring in English identified the VN collocations in these sentences. The manual VN collocations are compared with the instances extracted from the corpus and the result is showed in the Appendix. The evaluation indicates an average recall rate of 74.47% and precision of 66.67%.

Table 4
Experiment result of VN collocation extracted from Sinorama parallel Corpus

| #answer keys | #output | #Correct | Recall (%) | Precision (%) |
|--------------|---------|----------|------------|---------------|
| 94           | 105     | 70       | 74.47      | 66.67         |

It is very difficult to evaluation the experimental results. There were obvious and clear-cut collocations and non collocation, but there were a lot of cases such as “improve environment” and “share housework” that were difficult to judge and may be evaluated differently by different people. There is room for improvement as far as recall and precision ratios are concerned. Nevertheless, the extracted VNs are very diverse and useful for language learning purpose.

4 Discussion

The proposed approach offers a simple algorithm for automatic acquisition of the VN instances from a corpus. The method is particularly interested in following ways:

i. We use a data-driven approach to extract monolingual collocations.
ii. The algorithm is applicable to elastic collocations.
iii. Systematic way of setting thresholds for a certain level of confidence
iv. We could obtained instances of VN collocation through the simple statistical information.

While Xtract extracts VN types, we focus on the VN instances. It is understandable that we would get slightly lower recall and precision rates.

5 Conclusion & Future work

In this paper, we describe an algorithm that employs statistical analyses to extract instance of VN collocations from a corpus. The algorithm is applicable to elastic collocations. The main difference between our algorithm and Xtract lies in that we extract the instances from the sentence instead of extracting the VN types directly.

Moreover, in our research we observe other types related to VN such as VP (ie. verb + preposition) and VNP (ie. verb + noun + preposition). In the future, we will further take these two patterns into consideration to extract more types of verb-related collocations.

References

Benson, Morton., Evelyn Benson, and Robert Ilson. The BBI Combinatory Dictionary of English: A Guide to Word Combinations. John Benjamins, Amsterdam, Netherlands, 1986.
Choueka, Y. (1988) : "Looking for needles in a haystack", Actes RIAO, Conference on User-Oriented Context Based Text and Image Handling, Cambridge, p. 609-623.

Choueka, Y.; Klein, and Neuwitz, E.. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. Journal of the Association for Literary and Linguistic Computing, 4(1):34-8, (1983)

Church, K. W. and Hanks, P. Word association norms, mutual information, and lexicography. Computational Linguistics, 1990, 16(1), pp. 22-29.

Dagan, I. and K. Church. Termight: Identifying and translation technical terminology. In Proc. of the 4th Conference on Applied Natural Language Processing (ANLP), pages 34-40, Stuttgart, Germany, 1994.

Dunning, T (1993) Accurate methods for the statistics of surprise and coincidence, Computational Linguistics 19:1, 61-75.

Haruno, M., S. Ikehara, and T. Yamazaki. Learning bilingual collocations by word-level sorting. In Proc. of the 16th International Conference on Computational Linguistics (COLING '96), Copenhagen, Denmark, 1996.

Huang, C.-R., K.-J. Chen, Y.-Y. Yang, Character-based Collocation for Mandarin Chinese, In ACL 2000, 540-543.

Inkpen, Diana Zaiu and Hirst, Graeme. "Acquiring collocations for lexical choice between near-synonyms." SIGLEX Workshop on Unsupervised Lexical Acquisition, 40th meeting of the Association for Computational Lin

Justeson, J.S. and Slava M. Katz (1995). Technical Terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(1):9-27.

Kupiec, Julian. An algorithm for finding noun phrase correspondences in bilingual corpora. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 1993.

Lin, D. Using collocation statistics in information extraction. In Proc. of the Seventh Message Understanding Conference (MUC-7), 1998.

Melamed, I. Dan. "A Word-to-Word Model of Translational Equivalence". In Procs. of the ACL97. pp 490-497. Madrid Spain, 1997.

Smadja, F. 1993. Retrieving collocations from text: Xtract. Computational Linguistics, 19(1):143-177

Smadja, F., K.R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics, 22(1):1-38, 1996.

**Appendix**

The manual VN collocations are compared with the instances extracted from the corpus:

| Rec  | Manual VN collocations                  | Automatic extracting VN collocations                       |
|------|----------------------------------------|-----------------------------------------------------------|
| 162  | ask question \ hold conference \ grant amnesty \ realize probability | grant amnesty \ ask question \ realize probability         |
| 1647 | enforce rule (強化) \ break rule \ enhance image \ forge reputation \ respect law | enhance image \ respect law \ improve organization \ break rule \ reward reputation |
| 2106 |                                                          |                                                           |
| 4898 |                                                          |                                                           |
| 5857 | take power \ do research                  | done research \ accuse linguistics                        |
| 6489 | make demand \ make improvement \ make breakthrough | make breakthrough                                          |
| 6871 | put mark \ release album                  | release album \ put mark                                  |
| 6887 |                                                          |                                                           |
| 7420 | take risk \ make start                    | take risk \ make start \ lead risk                       |
| 7710 | make money \ make profit \ rise price | stop conglomerate \ make money \ rise price \ make profit |
| 7932 | eliminate unfairness \ seek equity | seek equity \ eliminate unfairness |
| 8056 | | |
| 8510 | improve environment | improve environment |
| 9326 | do research \ have influence | have influence |
| 9433 | | |
| 10600 | | |
| 10624 | contemplate footstep | |
| 11293 | understand meaning | understand meaning |
| 11603 | | |
| 12937 | receive attention \ witness progress | receive attention \ witness progress |
| 13033 | promote idea \ invest effort \ share housework \ expend effort | expend effort \ share housework \ promote idea \ invest effort |
| 13491 | | |
| 13576 | test wisdom | |
| 15349 | take paycut \ exceed budget \ unload property | show increase \ house price \ unload property |
| 16949 | | |
| 17106 | block view \ make offering | make offering |
| 17608 | lose ability | lose ability \ save forest |
| 17924 | take effort \ take time | consider success |
| 18183 | | |
| 18717 | carry work | carry work |
| 18745 | | |
| 19735 | bear son | bear son |
| 20002 | make money \ think way | make money \ think way |
| 21450 | buy portion | |
| 21663 | live life | live space |
| 22610 | | |
| 22306 | adopt method | adopt method |
| 23074 | | |
| 24307 | move production | move production \ develop computer |
| 25478 | make thing | make thing |
| 26030 | | |
| 28303 | increase chance \ increase production | increase chance \ increase production |
| 28336 | | |
| 28417 | write essay | write essay |
| 28806 | write seller | write seller |
| 28826 | | |
| 29003 | make money \ take care \ have time | take care \ make money \ have time |
| 29292 | | |
| 29736 | damage environment | damage environment \ insure recovery \ choose styrofoam \ recover styrofoam |
| 30881 | donate kidney \ implant kidney | donate kidney \ implant kidney |
| 31096 | drive car \ take transportation \ have responsibility | drive car \ consume pastry \ have responsibility \ wrap candy |
| 32975 | instruct student | instruct student |
| 33558 | take part in |
| 33993 | take part \ detail research |
| 33994 | have chance |
| 34008 | excite pupil |
| 34966 | have drink \ kick habit |
| 35113 | carry card \ ask carrier \ have drink \ kick habit |
| 35898 | come face |
| 35906 | announce approval \ bear child |
| 36931 | apply concept \ build contact |
| 36988 | apply concept |
| 37025 | start movement |
| 37811 | hear sound |
| 37916 | announcement approval \ bear child |
| 38197 | make adjustment \ build contact |
| 38200 | make adjustment |
| 38231 | spend day |
| 38626 | spend day |
| 40823 | do service |
| 40873 | put effort \ dedicate life |
| 41102 | dedicate life \ achieve dream \ put effort |
| 41383 | gain influence \ spend day |
| 41532 | spend day |
| 43027 | pay a visit to |
| 43199 | follow road |
| 43304 | derive satisfaction |
| 43465 | derive satisfaction |
| 44052 | strip circle |
| 44276 | impose sanction \ endanger specie |
| 44990 | pay a visit to |
| 45187 | pay a visit |
| 45499 | pay visit |
| 45756 | stope frame |
| 45857 | point way |
| 45905 | point way |
| 46466 | charge | strip circle |
| 47134 | offend policeman |
| 47226 | offend policeman |
| 47337 | borrow hairpin | offend policeman |
| 47720 | receive treatment |
| 48694 | elapse step |