TANGO: Bilingual Collocational Concordancer

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

Yu-Chia Chang
Inst. of Information System and Applictaion
National Tsing Hua University
101, Kuangfu Road, Hsinchu, Taiwan
u881222@alumni.nthu.edu.tw

Jason S. Chang
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, Taiwan
jschang@cs.nthu.edu.tw

Abstract

In this paper, we describe TANGO as a collocational concordancer for looking up collocations. The system was designed to answer user's query of bilingual collocational usage for nouns, verbs and adjectives. We first obtained collocations from the large monolingual *British National Corpus* (BNC). Subsequently, we identified collocation instances and translation counterparts in the bilingual corpus such as *Sinorama Parallel Corpus* (SPC) by exploiting the word-alignment technique. The main goal of the concordancer is to provide the user with a reference tools for correct collocation use so as to assist second language learners to acquire the most eminent characteristic of native-like writing.

1 Introduction

Collocations are a phenomenon of word combination occurring together relatively often. Collocations also reflect the speaker's fluency of a language, and serve as a hallmark of near native-like language capability.

Collocation extraction is critical to a range of studies and applications, including natural language generation, computer assisted language learning, machine translation, lexicography, word sense disambiguation, cross language information retrieval, and so on.

Hanks and Church (1990) proposed using point-wise mutual information to identify collocations in lexicography; however, the method may result in unacceptable collocations for low-count pairs. The best methods for extracting collocations usually take into consideration both linguistic and statistical constraints. Smadja (1993) also detailed techniques for collocation extraction and developed a program called XTRACT, which is capable of computing flexible collocations based on elaborated statistical calculation. Moreover, log likelihood ratios are regarded as a more effective method to identify collocations especially when the occurrence count is very low (Dunning, 1993).

Smadja's XTRACT is the pioneering work on extracting collocation types. XTRACT employed three different statistical measures related to how associated a pair to be collocation type. It is complicated to set different thresholds for each statistical measure. We decided to research and develop a new and simple method to extract monolingual collocations.

We also provide a web-based user interface capable of searching those collocations and its usage. The concordancer supports language learners to acquire the usage of collocation. In the following section, we give a brief overview of the TANGO concordancer.

2 TANGO

TANGO is a concordancer capable of answering users' queries on collocation use. Currently, TANGO supports two text collections: a monolingual corpus (BNC) and a bilingual corpus (SPC). The system consists of four main parts:

2.1 Chunk and Clause Information Integrated

For CoNLL-2000 shared task, chunking is considered as a process that divides a sentence into syntactically correlated parts of words. With the benefits of CoNLL training data, we built a chunker that turn sentences into smaller syntactic structure of non-recursive basic phrases to facilitate precise collocation extraction. It becomes easier to identify the argument-predicate relationship by looking at adjacent chunks. By doing so, we save time as opposed to n-gram statistics or full parsing. Take a text in CoNLL-2000 for example:

The words correlated with the same chunk tag can be further grouped together (see Table 1). For instance, with chunk information, we can extract

```
  Confidence/B-NP    in/B-PP    the/B-NP
pound/I-NP   is/B-VP   widely/I-VP   ex-
pected/I-VP   to/I-VP   take/I-VP   an-
other/B-NP  sharp/I-NP  dive/I-NP  if/B-
SBAR  trade/B-NP  figures/I-NP  for/B-PP
September/B-NP
```

(Note: Every chunk type is associated with two different chunk tags: B-CHUNK for the first word of the chunk and I-CHUNK for the other words in the same chunk)

the target VN collocation "take dive" from the example by considering the last word of two adjacent VP and NP chunks. We build a robust and efficient chunking model from training data of the CoNLL shared task, with up to 93.7% precision and recall.

Sentence chunking	Features
Confidence	NP
in	PP
the pound	NP
is expected to take	VP
another sharp dive	NP
if	SBAR
trade figures	NP
for	PP
September	NP

Table 1: Chunked Sentence

In some cases, only considering the chunk information is not enough. For example, the sentence "…the attitude he had towards the country is positive…" may cause problem. With the chunk information, the system extracts out the type "have towards the country" as a VPN collocation, yet that obviously cuts across two clauses and is not a valid collocation. To avoid that kind of errors, we further take the clause information into account.

With the training and test data from CoNLL-2001, we built an efficient HMM model to identify clause relation between words. The language model provides sufficient information to avoid extracting wrong collocations. Examples show as follows (additional clause tags will be attached):

(1) ….the attitude *(S** he has ***S)* toward the country

(2) *(S** I think *(S** that the people are most concerned with the question of *(S** when conditions may become ripe. ***S)S)S)*

As a result, we can avoid combining a verb with an irrelevant noun as its collocate as "*have toward country*" in (1) or "*think … people*" in (2). When the sentences in the corpus are annotated with the chunk and clause information, we can consequently extract collocations more precisely.

2.2 Collocation Type Extraction

A large set of collocation candidates can be obtained from BNC, via the process of integrating chunk and clause information. We here consider three prevalent Verb-Noun collocation structures in corpus: VP+NP, VP+PP+NP, and VP+NP+PP. Exploiting Logarithmic Likelihood Ratio (LLR) statistics, we can calculate the strength of association between two collocates. The collocational type with threshold higher than 7.88 (confidence level 99.5%) will be kept as one entry in our collocation type list.

2.3 Collocation Instance Identification

We subsequently identify collocation instances in the bilingual corpus (SPC) with the collocation types extracted from BNC in the previous step. Making use of the sequence of chunk types, we again single out the adjacent structures of VN, VPN, and VNP. With the help of chunk and clause information, we thus find the valid instances where the expected collocation types are located, so as to build a collocational concordance. Moreover, the quantity and quality of BNC also facilitate the collocation identification in another smaller bilingual corpus with better statistic measure.

English sentence	Chinese sentence
If in this time no one shows concern for them, and directs them to correct thinking, and teaches them how to express and release emotions, this could very easily leave them with a terrible personality complex they can never resolve.	如果這時沒有人關心他們，引導他們正確思考，教他們表達、宣洩**情緒**，極易在人格成長上留下一個打不開的死結。
Occasionally some kungfu movies may appeal to foreign audiences, but these too are exceptions to the rule.	偶爾有一些武打片對某些外國**觀眾**有吸引力，但也是個案。

Table 2: Examples of collocational translation memory

Type	Collocation types in BNC
VN	631,638
VPN	15,394
VNP	14,008

Table 3: The result of collocation types extracted from BNC and collocation instances identified in SPC

2.4 Extracting Collocational Translation Equivalents in Bilingual Corpus

When accurate instances are obtained from bilingual corpus, we continue to integrate the statistical word-alignment techniques (Melamed, 1997) and dictionaries to find the translation candidates for each of the two collocates. We first locate the translation of the noun. Subsequently, we locate the verb nearest to the noun translation to find the translation for the verb. We can think of collocation with corresponding translations as a kind of translation memory (shows in Table 2).The implementation result of BNC and SPC shows in the Table 3, 4, and 5.

3 Collocation Concordance

With the collocation types and instances extracted from the corpus, we built an online collocational concordancer called TANGO for looking up translation memory. A user can type in any English query and select the intended part of speech of query and collocate. For example in Figure 1, after query for the verb collocates of the noun "influence" is submitted, the results are displayed on the return page. The user can then browse through different collocates types and also click to get to see all the instances of a certain collocation type.

Noun	VN types
Language	320
Influence	319
Threat	222
Doubt	199
Crime	183
Phone	137
Cigarette	121
Throat	86
Living	79
Suicide	47

Table 4: Examples of collocation types including a given noun in BNC

VN type	Example
Exert influence	That means they would already be **exerting** their **influence** by the time the microwave background was born.
Exercise influence	The Davies brothers, Adrian (who scored 14 points) and Graham (four), **exercised** an important creative **influence** on Cambridge fortunes while their flankers Holmes and Pool-Jones were full of fire and tenacity in the loose.
Wield influence	Fortunately, George V had worked well with his father and knew the nature of the current political trends, but he did not **wield** the same **influence** internationally as his esteemed father.

Table 5: Examples of collocation instances extracted from SPC

Moreover, using the technique of bilingual collocation alignment and sentence alignment, the system will display the target collocation with highlight to show translation equivalents in context. Translators or learners, through this web-based interface, can easily acquire the usage of each collocation with relevant instances. This collocational concordancer is a very useful tool for self-inductive learning tailored to intermedi-ate or advanced English learners.

Users can obtain the result of the VN or AN collocations related to their query. TANGO shows the collocation types and instances with collocations and translation counterparts highlighted.

The evaluation (shows in Table 6) indicates an average precision of 89.3 % with regard to satisfactory.

4 Conclusion and Future Work

In this paper, we describe an algorithm that employs linguistic and statistical analyses to extract instance of VN collocations from a very large corpus; we also identify the corresponding translations in a parallel corpus. The algorithm is applicable to other types of collocations without being limited by collocation's span. The main difference between our algorithm and previous work lies in that we extract valid instances instead of types, based on linguistic information of chunks and clauses. Moreover, in our research we observe

Type	The number of selected sentences	Translation Memory	Translation Memory (*)	Precision of Translation Memory	Precision of Translation Memory (*)
VN	100	73	90	73	90
VPN	100	66	89	66	89
VNP	100	78	89	78	89

Table 6: Experiment result of collocational translation memory from Sinorama parallel Corpus

Figure 1: The caption of the table

other types related to VN such as VPN (ie. verb + preposition + noun) and VNP (ie. verb + noun + preposition), which will also be crucial for machine translation and computer assisted language learning. In the future, we will apply our method to more types of collocations, to pave the way for more comprehensive applications.

Acknowledgements

This work is carried out under the project "CANDLE" funded by National Science Council in Taiwan (NSC92-2524-S007-002). Further information about CANDLE is available at http://candle.cs.nthu.edu.tw/.

References

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

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

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.