Collocational Translation Memory Extraction Based on Statistical and Linguistic Information

Jia-Yan Jian
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, Taiwan
d914339@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 propose a new method for bilingual collocation extraction from a parallel corpus to provide phrasal translation memory. The method integrates statistical and linguistic information for effective extraction of collocations. The linguistic information includes parts of speech, chunks, and clauses. With an implementation of the method, we obtain first an extended list of collocations from monolingual corpora such as British National Corpus (BNC). Subsequently, we exploit the list to identify English collocations in Sinorama Parallel Corpus (SPC). Finally, we use word alignment techniques to retrieve the translation equivalent of English collocations from the bilingual corpus, so as to provide phrasal translation memory for machine translation system. Based on the strength of chunk and clause analyses, we are able to extract a large number of collocations and translations with much less time and effort than those required by N-gram analysis or full parsing. Furthermore, we also consider longer collocation pattern such as a preposition involved in VN collocation. In the future, we plan to extend the method to other types of collocation.

Keyword. Bilingual Collocation Extraction, Collocational Translation Memory, Collocational Concordancer

1 Introduction

Example-based machine translation (EBMT), a corpus-based MT method, has been recently suggested as an efficient step toward automatic translation (Nagao, 1981; Kitano, 1993, Carl, 1999, Andrimanankasian et al., 1999; Brown, 2000). Under the approach, systems exploited examples similar to input and adjusted the translations to obtain the result. Translations are preprocessed and stored in a translation memory which serves as an archive of existing translation for MT system to reuse. Nowadays, there have been a number of transducers applied to convert sentences in bilingual corpus into translation patterns, which can be further exploited as a translation memory, such as Transit[1], Deja–Vu[2], TransSearch[3], TOTALrecall[4], and so on.

A problem that most MT system may encounter is the collocational translation if the system intends not to literally translate the input text. This smaller syntax unit not only facilitates a more native-like translation, but also enhances the performance of recent EBMT system. Elastic collocation structure provides more flexibility in handle translation pattern as in "…not yet to **take** what he wants **into consideration**…"

---

[1] *Transit* (http://www.star-group.net/eng/software/sprachtech/transit.html)
[2] *Deja–Vu* (http://www.atril.com/)
[3] *TransSearch* (http://www.tsrali.com/)
[4] *TOTALrecall*(http://candle.cs.nthu.edu.tw/Counter/Counter.asp?funcID=1)

2 Extraction of Collocational Translation Memory

Using valuable linguistic information—chunk and clause analyses, we can retrieve Verb-Noun collocations from a large corpus (i.e. BNC) with good quality and quantity. We further use this collocation type list to identify the concise collocational instances in a bilingual corpus (i.e. SPC). We also use word-alignment technique to extract the matching translation of verb and noun respectively, so as to obtain phrasal translation memory. The detailed approach is described in this section:

2.1 Chunk and Clause Information Integrated

CoNLL-2000[5] shared task considered text chunking as a process that divides a text into syntactically correlated parts of words. With the benefits of chunk information, we can chunk the sentence into smaller syntactic structure which facilitates precise collocation extraction. It becomes easier to identify the argument-predicate relationship between each chunk, and save more time to extract as opposed to full parsing. Take a passage in CoNLL-2000 benchmark for example:

Confidence/B-NP in/B-PP the/B-NP pound/I-NP is/B-VP
widely/I-VP expected/I-VP to/I-VP take/I-VP another/B-NP
sharp/I-NP dive/I-NP if/B-SBAR trade/B-NP figures/I-NP for/B-PP
September/B-NP

Note: I-NP for noun phrase words and I-VP for verb phrase words. Most chunk types have 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 words in the same chunk can be further grouped together (as in Table 1). With chunk information, we can extract the target VN collocation, "take … dive" from the text by considering the last word of each adjacent VP and NP chunks. We built a robust and efficient chunker from the training data of the CoNLL shared task, with over 93% precision and recall[6].

Table 1: Chunked Sentence

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

[5] CoNLL is the yearly meeting of the SIGNLL, the Special Interest Group on Natural Language Learning of the Association for Computational Linguistics. The shared task of text chunking in CoNLL-2000 is available at http://cnts.uia.ac.be/conll2000/.

[6] We built the chunker from shared CoNLL-2000 training data and evaluate the result with the test data provided by CoNLL-2000. The precision and the recall are both 93.7%.

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 VP + PP + NP, yet this one is erroneous because it cuts across two clauses. To avoid this case, we further take the clause information into account.

With the training data from CoNLL-2001, we built an efficient clause model based on HMM to identify the clause relation between words. The language model provides sufficient information to avoid extracting wrong VN collocation instances. 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 the verb from being combined with the irrelevant noun as its collocate (as in (1)) or extracting the adjacent noun serving as the subject of another clause (as in (2)). When the sentences in the corpus are preprocessed with the chunk and clause identification, we can consequently assure high accuracy of collocation extraction.

Log-likelihood ratio : LLR(x;y)

$$LLR(x;y) = -2\log_2 \frac{p_1^{k_1}(1-p_1)^{n_1-k_1}(1-p_2)^{n_2-k_2}}{p^{k_1}(1-p)^{n_1-k_1}\ p^{k_2}(1-p)^{n_2-k_2}}$$

k_1 : of pairs that contain x and y simultaneously.
k_2 : of pairs that contain x but do not contain y.
n_1 : of pairs that contain y
n_2 : of pairs that does not contain y
$p_1 = k_1 / n_1$
$p_2 = k_2 / n_2$
$p = (k_1+k_2) / (n_1+n_2)$

2.2 Extraction of Collocation Types

A huge 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 each two collocates. The collocational type with threshold higher than 7.88 (confidence level 95%) will serve as one entry in our collocation type list.

2.3 Extraction of Collocation Instances

We subsequently identify collocation instances in the Sinorama Parallel Corpus (SPC) matching the collocation types extracted from BNC. Making use of the sequence of chunk types, we again single out the adjacent structures: VP+NP, VP+PP+NP, or VP+NP+PP. 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.

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).

Table 2: Examples of collocational translation memory

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、渲洩**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。

3 Implementation and evaluation

We extracted VN collocations from the BNC which contains about 4 million sentences, and obtained 631,638 VN, 15,394 VPN, and 14,008 VNP collocation types with an implementation of the proposed method. We continued to identify 26,315VN, 3,457 VPN, and 4,406 VNP collocation instances in SPC and generated eligible translation memory via word-alignment techniques. The implementation result of BNC and SPC shows in the Table 3, 4, and 5.

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

Type	Collocation types in British Nation Corpus (BNC)	Collocation instances in Sinorama Parallel Corpus (SPC)
VN	631,638	26,315
VPN	15,394	3,457
VNP	14,008	4,406

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

Noun	VN types	VN instances
Language	320	945
Influence	319	880
Threat	222	633
Doubt	199	545
Crime	183	498
Phone	137	460
Cigarette	121	379
Throat	86	246
Living	79	220
Suicide	47	134

Table 5:Examples of collocation instances extracted from SPC

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.
Extend influence	The cab extended its influence into the non-government sector, funding research by the Cathedral Advisory Commission and the Royal Society for the Protection of Birds.
Reflect influence	The general standard of farming was good, reflecting the influence of the sons who had attained either a degree or a diploma in agriculture before returning home.
Diminish influence	To break up the Union now would diminish our influence for good in the world, just at the time when it is most needed.
Gain influence	In general, women have not benefited much in the job market from capitalist industrialization nor have they gained much influence in society outside the family through political channels.
Counteract influence	To try and counteract the influence of the extremists, the moderate wing of the party launched a Labour Solidarity Campaign in 1981.
Reduce influence	Whether the curbs on police investigation will reduce police influence on the outcome of the criminal process is not easy to determine.
Show influence	Ellis and Shepherd (1974) first drew attention to this but a number of experiments by Young and his colleagues have failed to show any influence of age of acquisition of words on dichotic listening (Young and Ellis , 1980) or tachistoscopic hemifield asymmetry (Ellis and Young , 1977 ; Young and Bion , 1980b) even when it is the age at which words are first read rather than heard that is under investigation.

As for each collocation type, we randomly selected 100 test sentences for manual evaluation. A human judge, who majored in Foreign Languages, assessed the result of the matching translation. The evaluation was done by judging whether the corresponding collocational translation is valid or not. The three levels of quality were set: satisfactory translation, approximant translation (partial matching), and unacceptable translation. The examples of each level are shown in Table 6.

Table 6: Three levels of quality of the extracted translation memory

Level of quality	English sentences	Chinese sentences
satisfactory translation	Thus when Chinpao Shan put out its advertisement last year, looking for new people to develop its related **enterprises**, the notice frankly stated "Southern Taiwanese preferred.	去年，金寶山在發展**關係企業**徵招新人的廣告上，就坦白指明「本省籍南部人優先」。
approximant translation	Ah-ying relates that "Teacher Chang" friendly and easy-going, is always there to answer her **questions**. She even goes to him for answers when her friends have legal questions.	阿英表示，「張老師」親切隨和，只要有不懂的事，都去問老師，就連朋友有法律上的**問題**，也去請教他。
unacceptable translation	Said one observer, "If I can speak bluntly, the mainlanders are robbing graves of their treasures and smuggling them away, and the situation is bad. In reality, though, it is Taiwan that is behind it all committing the **crime**.	「說得不好聽，大陸近年來盜墓、文物走私情形嚴重，台灣其實是背後的劊子手！」有人這樣**認為**。

The evaluation result indicates an average precision rate of 89 % with regard to both satisfactory and approximant translation memory (shows in Table 7).

Table 7: Experiment result of collocational translation memory from Sinorama parallel Corpus

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

The average precision of translation memory: 72.3%
The average precision of translation memory (*): 89.3%

(*) stands for the numbers of translation memory which includes approximant translation.

4. Discussion and limitation

Collocation, a hallmark of near native speaker, is an important area in translation yet has long been neglected. Traditional machine translation tends to translate input texts word by word, which easily leads to literal translation. Therefore, even with abundant vocabulary from dictionary and grammar rule-based model systems still fail to generate fluent translation into a target language. For example, with the lack of collocational knowledge, machine translation system may recognize take as “na” (i.e. take away) and medicine as “yao” (i.e. medicine) in Chinese respectively. Thus, systems are inclined to literally translate take medicine into “na yao” (i.e. take away the medicine), and probably result in odd translation or mistranslation. We suggest that machine translation system take collocational translation memory into consideration for improved translation quality. The notion of collocation is also consistent with Example-Based Machine Translation (EBMT).

Due to the limitation of word-alignment technique, our method may incorrectly recognize some matching translation. We need better word-alignment to align translations more correctly. Moreover, the expansion of bilingual corpora can also increase the precision of retrieving collocational translation memory. It enables us to obtain enough counts for each collocate (i.e. verb and noun in VN collocation) in the target language so as to increase the reliability with the LLR statistics, which in turn eradicates the anomalous collocational translation memory.

5. Application: Collocational Concordance—TANGO

With the collocation types and instances extracted from the corpus, we built an on-line collocational concordance called TANGO for looking up collocation instances and translations. A user can type in any English words as query and select the expected part of speech of the accompanying words. For example in Figure 1, after query “influence” is submitted, the result of possible collocates will be displayed on the return page. The user can even select different adjacent collocates for further investigation. 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 bilingual collocational concordance is a very useful tool for self-inductive learning tailored to intermediate or advanced English learners.

Figure 1 Web-based Collocational Concordance

6. Conclusion

In the field of the machine translation, the Example-Based Machine Translation (EBMT) exploits existing translations in the hope of producing better quality in translation. However, the importance of collocational translation has always been neglected and hard to be dealt with. We propose the collocational translation memory — to provide a better translation method, intending to solve some problem encountered by literal translation. With satisfactory precision rates of collocation and translation extraction, we hope collocational translation memory will path ways to more applications in translation and computer assisted language learning.

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/.

Reference

[1] Andriamanankasina, T., Araki, K. and Tochinai, T. 1999. Example-Based Machine Translation of Part-Of-Speech Tagged Sentences by Recursive Division. Proceedings of MT SUMMIT VII. Singapore.

[2] Brown, R. D. 2000. Automated Generalization of Translation Examples. In Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000), pp. 125-131. Saarbrücken, Germany, August 2000.

[3] Carl, M. 1999. Inducing Translation Templates for Example-Based Machine Translation, Proc. of MT Summit VII.

[4] Melamed, I. D. 1997. A Word-to-Word Model of Translational Equivalence. Proc. of the ACL97. pp 490-497. Madrid Spain, 1997.

[5] Kitano, H. 1993. A Comprehensive and PracticalModel of Memory-Based Machine Translation. Proc.of IJCAI-93. pp. 1276-1282.

[6] Nagao, M. 1981. A Framework of a MechanicalTranslation between Japanese and English byAnalogy Principle, in Artificial and Human Intelligence, A. Elithorn and R. Banerji (eds.) North-Holland, pp. 173-180, 1984.