

Chinese Sketch Engine and the Extraction of Grammatical Collocations

Chu-Ren Huang	Adam Kilgarriff	Yiching Wu	Chih-Ming Chiu
Inst. of Linguistics	Lexicography MasterClass	Inst. of Linguistics	Inst. of Information Science
Academia Sinica	Information Technology	Tsing Hua University	Academia Sinica
churen@sinica.edu.tw	adam@lexmasterclass.com	d898702@oz.nthu.edu.tw	henning@hp.iis.sinica.edu.tw
Simon Smith	Pavel Rychly	Ming-Hong Bai	Keh-Jiann Chen
Dept. of Applied English	Faculty of Informatics	Inst. of Information Science	Inst. of Information Science
Ming Chuan University	Masaryk University.	Academia Sinica	Academia Sinica
ssmith@mcu.edu.tw	pary@textforge.cz	mhbai@sinica.edu.tw	kchen@iis.sinica.edu.tw

Abstract. This paper introduces a new technology for collocation extraction in Chinese. Sketch Engine (Kilgarriff et al., 2004) has proven to be a very effective tool for automatic description of lexical information, including collocation extraction, based on large-scale corpus. The original work of Sketch Engine was based on BNC. We extend Sketch Engine to Chinese based on Gigaword corpus from LDC. We discuss the available functions of the prototype Chinese Sketch Engine (CSE) as well as the robustness of language-independent adaptation of Sketch Engine. We conclude by discussing how Chinese-specific linguistic information can be incorporated to improve the CSE prototype.

1. Introduction

The accessibility to large scale corpora, at one billion words or above, has become both a blessing and a challenge for NLP research. How to efficiently use a gargantuan corpus is an urgent issue concerned by both users and corpora designers. Adam Kilgarriff et al. (2004) developed the Sketch Engine to facilitate efficient use of corpora. Their claims are two folded: that genuine linguistic generalizations

can be automatically extracted from a corpus with simple collocation information provided that the corpus is large enough; and that such a methodology is easily adaptable for a new language. The first claim was fully substantiated with their work on BNC. The current paper deals with the second claim by adapting the Sketch Engine to Chinese.

2. Online Chinese Corpora: The State of the Arts

2.1 Chinese Corpora

The first online tagged Chinese corpus is Academia Sinica Balanced Corpus of Modern Chinese (Sinica Corpus), which has been web-accessible since November, 1996. The current version contains 5.2028 million words (7.8927 million characters). The corpus data was collected between 1990 and 1996 (CKIP, 1995/1998). Two additional Chinese corpora were made available on line in 2003. The first is the Sinorama Chinese-English Parallel Text Corpus (Sinorama Corpus). The Sinorama Corpus is composed of 2,373 parallel texts in both Chinese and English that were published between 1976 and 2000. There are 103,252 pairs of sentences, composed of roughly 3.2 million

English words and 5.3 million Chinese characters¹. The second one is the modern Chinese corpus developed by the Center for Chinese Linguistics (CCL Corpus) at Peking University. It contains eighty-five million (85,398,433) simplified Chinese characters which were published after 1919 A.D.

2.2 Extracting Linguistic Information from Online Chinese Corpora: Tools and Interfaces

The Chinese corpora discussed above are all equipped with an online interface to allow users to extract linguistic generalizations. Both Sinica Corpus and CCL Corpus offer KWIC-based functions, while Sinorama Corpus gives sentence and paragraph aligned output.

2.2.1 String Matching or Word Matching

The basic unit of query that a corpus allows defines the set of information that can be extracted from that corpus. While there is no doubt that segmented corpus allows more precise linguistic generalizations, string-based collocation still afford a corpus of the robustness that is not restricted by an arbitrary word-list or segmentation algorithm. This robustness is of greatest value when extracting neologism or sub-lexical collocations. Since CCL Corpus is not segmented and tagged, string-based KWIC is its main tool for extracting generalizations. This comes with the familiar pitfall of word boundary ambiguity. For instance, a query of *ci.yao* 次要 ‘secondary’ may yield the intended result (1a), as well as noise (1b).

1a. 但这是次要的

dan zhe shi ci.yao de

but this is secondary DE

‘But this is secondary’

b. 他几次要求她答复

ta ji ci.yao.qiu ta da.fu

he several time ask her answer

‘He had asked her to answer for several times’

Sinica Corpus, on the other hand, is fully segmented and allows word-based generalizations. In addition, Sinica Corpus also allows wildcards in its search. Users specify a wildcard of arbitrary length (*), or fixed length (?). This allows search of a class of words sharing some character strings.

2.2.2 Display of Extracted Data

Formal restriction on the display of extracted data also constraints the type of information that can be obtained from that corpus. Sinica Corpus allows users to change window size from about 25 to 57 Chinese characters. However, since a Chinese sentence may be longer than 57 characters, Sinica Corpus cannot guarantee that a full sentence is displayed. CCL Corpus, on the other hand, is able to show a full output sentence, which may be up to 200 Chinese characters. However, it does not display more than a full sentence. Thus it cannot show discourse information. Sinorama Corpus with TOTALrecall interface is most versatile in this respect. Aligned bilingual full sentences are shown with an easy link to the full text.

In terms of size and completeness of extracted data, Sinica Corpus returns all matched examples. However, cut and paste must be performed for the user to build his/her dataset. CCL Corpus, on the other hand, limits data to 500 lines per page, but allows easy download of output data. Lastly, Sinorama/TOTALrecall provides choices of 5 to 100 sentences per page.

¹http://cio.nist.gov/esd/emaildir/lists/mt_list/msg00033.html

2.2.3 Refining Extracted Information: Filter and Sorter

Both Sinica Corpus and CCL corpus allows users to process extracted information, using linguistic and contextual filter or sorter. The CCL corpus requires users to remember the rules, while Sinica Corpus allows users to fill in blanks and/or choose from pull-down menu. In particular, Sinica Corpus allows users to refine their generalization by quantitatively characterizing the left and right contexts. The quantitative sorting functions allowed include both word and POS frequency, as well as word mutual information.

2.2.4 Extracting Grammatical Information

Availability of grammatical information depends on corpus annotation. CCL and Sinorama Corpus do not have POS tags. Sinica Corpus is the only Chinese corpus allowing users to access an overview of a keyword's syntactic behavior. Users can obtain a list of types and distribution of the keyword's syntactic category. In addition, users can find possible collocations of the keyword from the output of Mutual Information (MI).

The most salient grammatical information, such as grammatical functions (subject, object, adjunct etc.) is beyond the scope of the traditional corpus interface tools. Traditional corpora rely on the human users to arrive at these kinds of generalizations.

3. Sketch Engine: A New Corpus-based approach to Grammatical Information

Several existing linguistically annotated corpus of Chinese, e.g. Penn Chinese Tree Bank (Xia et al., 2000), Sinica Treebank (Chen et al., 2003), Proposition Bank (Xue and Palmer, 2003, 2005) and Mandarin VerbNet (Wu and Liu,

2003), suffer from the same problem. They are all extremely labor-intensive to build and typically have a narrow coverage. In addition, since structural assignment is theory-dependent and abstract, inter-annotator consistency is difficult to achieve. Since there is also no general consensus on the annotation scheme in Chinese NLP and linguistics, building an effective interface for public use is almost impossible.

The Sketch Engine offers an answer to the above issues.

3.1 Initial Implementation and Design of the Sketch Engine

The Sketch Engine is a corpus processing system developed in 2002 (Kilgarriff and Tugwell, 2002; Kilgarriff et al., 2004). The main components of the Sketch Engine are KWIC concordances, word sketches, grammatical relations, and a distributional thesaurus. In its first implementation, it takes as input basic BNC (British National Corpus, (Leech, 1992)) data: the annotated corpus, as well as list of lemmas with frequencies. In other words, the Sketch Engine has a relatively low threshold for the complexity of input corpus.

The Sketch Engine has a versatile query system. Users can restrict their query in any sub-corpus of BNC. A query string may be a word (with or without POS specification), or a phrasal segment. A query can also be performed using Corpus Query Language (CQL). The output display format can be adjusted, and the displayed window of a specific item can be freely expanded left and right. Most of all, the Sketch Engine produces a Word Sketch (Kilgarriff and Tugwell, 2002) that is an automatically generated grammatical description of a lemma in terms of corpus collocations. All items in each collocation are linked back to the original corpus data. Hence it is similar to a

Linguistic Knowledge Net anchored by a lexicon (Huang et al., 2001).

A Word Sketch is a one-page list of a keyword's functional distribution and collocation in the corpus. The functional distribution includes: subject, object, prepositional object, and modifier. Its collocations are described by a list of linguistically significant patterns in the language. Word Sketch uses regular expressions over POS-tags to formalize rules of collocation patterns, e.g. (2) is used to retrieve the verb-object relation in English:

2. 1:"V" "(DET|NUM|ADJ|ADV|N)"* 2:"N"

The expression in (2) says: extract the data containing a verb followed by a noun regardless of how many determiners, numerals, adjectives, adverbs and nouns preceding the noun. It can extract data containing *cook meals* and *cooking a five-course gala dinner*, and *cooked the/his/two surprisingly good meals* etc.

The Sketch Engine also produces thesaurus lists, for an adjective, a noun or a verb, the other words most similar to it in their use in the language (Kilgarriff et al. 2004). For instance, the top five synonym candidates for the verb *kill* are *shoot* (0.249), *murder* (0.23), *injure* (0.229), *attack* (0.223), and *die* (0.212).² It also provides direct links to the Sketch Difference which lists the similar and different patterns between a keyword and its similar word. For example, both *kill* and *murder* can occur with objects such as *people* and *wife*, but *murder* usually occurs with personal proper names and seldom selects animal nouns as complement whereas *kill* can take *fox*, *whale*, *dolphin*, and *guerrilla*, etc. as its object.

The Sketch Engine adopts Mutual

Information (MI) to measure the salience of a collocation. Salience data are shown against each collocation in Word Sketches and other Sketch Engine output. MI provides a measure of the degree of association of a given segment with others. Pointwise MI, calculated by Equation 3, is what is used in lexical processing to return the degree of association of two words x and y (a **collocation**).

$$3. I(x; y) = \log \frac{P(x|y)}{P(x)}$$

3.2 Application to Chinese Corpus

In order to show the cross-lingual robustness of the Sketch Engine as well as to propose a powerful tool for collocation extraction based on a large scale corpus with minimal pre-processing; we constructed Chinese Sketch Engine (CSE) by loading the Chinese Gigaword to the Sketch Engine (Kilgarriff et al., 2005). The Chinese Gigaword contains about 1.12 billion Chinese characters, including 735 million characters from Taiwan's Central News Agency, and 380 million characters from China's Xinhua News Agency³. Before loading Chinese Gigaword into Sketch Engine, all the simplified characters were converted into traditional characters, and the texts were segmented and POS tagged using the Academia Sinica segmentation and tagging system (Huang et al., 1997). An array of machine was used to process the 1.12 million characters, which took over 3 days to perform. All components of the Sketch Engine were implemented, including Concordance, Word Sketch, Thesaurus and Sketch Difference.

In our initial in-house testing of this prototype of the Chinese Sketch Engine, it does

² The similarity is measured and ranked adopting Lin's (1998) mathematics.

³<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T09>

produce the expected results with an easy to use interface. For instance, the Chinese Word Sketch correctly shows that the most common and salient object of *dai.bu* 逮捕 ‘to arrest’ is *xian.fan* 嫌犯 ‘suspect’; the most common subject *jing.fang* 警方 ‘police’; and the most common modifier *dang.chang* 當場.

The output data of Thesaurus correctly verify the following set of synonyms from the Chinese VerbNet Project: that *ren.wei* 認為 ‘to think’ behaves most like *biao.shi* 表示 ‘to express, to state’ (saliency 0.451), while *yi.wei* 以為 ‘to take somebody/something as’ is more like *jue.de* 覺得 ‘to feel, think’ (saliency 0.488). The synonymous relation can be illustrated by (4) and (5).

- 4a. 他認為到海外投資有一個觀念很重要，就是要知道當地的遊戲規則，接受這些條件。

ta ren.wei dao hai.wai tou.zi you yi ge guan.nian hen zhong.yao, jiu shi yao zhi.dao dang.di de you.xi gui.ze

‘He believes that for those investing overseas, there is a very important principle-one must know the local rules of the game, and accept them.’

- b. 執政黨也表示，由於公視爭議太大，恐怕無法全力支持。

zhi.zheng.dang ye biao.shi, you.yu gong.shi zheng.yi tai da, kong.pa wu.fa quan.li zhi.chi

‘The KMT also commented that due to the many controversies surrounding PTV, it could not wholeheartedly support it either.’

- 5a. 何家駒就認為：「電視有基本語言和文法，要講究賣點和市場。」

he.jia.ju jiu ren.wei : 「dian.shi you ji.ben yu.yan he wen.fa, yao jiang.jiu mai.dian he shi.chang。」

‘Ho Chia-chu says, "Television has its own fundamental language and grammar. You must consider selling points and the market."’

- b. 她表示：「我希望佛教徒能瞭解，父權社會與覺悟的社會是不相合的。」

ta biao.shi : 「wo xi.wang fuo.jiao.tu neng liao.jie, fu.quan she.hui yu jue.wu de she.hui shi bu xiang.he de。」

‘She says "I hope that followers of Buddhism can realize that a patriarchal society is incompatible with an enlightened society."’

The above examples show that *ren.wei* and *biao.shi* can take both direct and indirect quotation. *Yi.wei* and *jue.de*, on the other hand, can only be used in reportage and cannot introduce direct quotation.

Distinction between near synonymous pairs can be obtained from Sketch Difference. This function is verified with results from Tsai et al.’s study on *gao.xing* 高興 ‘glad’ and *kuai.le* 快樂 ‘happy’ (Tsai et al., 1998). *Gao.xing* ‘glad’ specific patterns include the negative imperative *bie* 別 ‘don’t’. It also has a dominant collocation with the potentiality complement marker *de* 得 (e.g. *ta gao.xing de you jiao you tiao* 她高興得又叫又跳 ‘she was so happy that she cried and danced’). In contrast, *kuai.le* ‘happy’ has the specific collocation with holiday nouns such as *qiu.jie* 秋節 ‘Autumn Festival’. The Sketch Difference result is consistent with the account that *gao.xing/kuai.le* contrast is that inchoative state vs. homogeneous state.

4. Evaluation and Future Developments

An important feature of the prototype of the Chinese Sketch Engine is that, in order to test the robustness of the Sketch Engine design, the original regular expression patterns were adopted with minimal modification for Chinese. Even though both are SVO languages with similar surface word order, it is obvious that they differ substantially in terms of assignment of grammatical functions. In addition, the Sinica tagset is different from the BNC tagset and

actually has much richer functional information. These are the two main directions that we will pursue in modification and improvement of the Chinese Sketch Engine.

4.1 Word Boundary Representation

Word breaks are not conventionalized in Chinese texts. This poses a challenge in Chinese language processing. The Chinese Sketch Engine inserted space after segmentation, which helps to visualize words. In the future, it will be trivial to allow the conventional alternative of no word boundary markups. However, it will not be trivial to implement fuzzy function to allow searches for non-canonical lemmas (i.e. lemmas that are segmented differently from the standard corpus).

4.2 Sub-Corpora Comparison

The Chinese Gigaword corpus is marked with two different genres, story and non-story. A still more salient sub-corpus demarcation is the one between Mainland China corpus and Taiwan corpus. Sketch Difference between lemmas from two sub-corpora is being planned. This would allow future comparative studies and would have wide applications in the localization adaptations of language related applications.

4.3 Collating Frequency Information with POS

One of the convenient features of Sketch Engine that a frequency ranked word list is linked to all major components. This allows a very easy and informative reference. Since cross-categorical derivation with zero morphology is dominant in Chinese, it would help the processing greatly if POS information is added to the word list. Adding such information would also open the possibility of accessing the POS ranked frequency information.

4.5 Fine-tuning Collocation Patterns

The Sketch Engine relies on collocation patterns, such as (2) above, to extract collocations. The regular expression format allows fast processing of large scale corpora with good results. However, these patterns can be fine-tuned for better results. We give VN collocates with object function as example here. In (6), verbs are underlined with a single line, and the collocated nouns identified by English Word Sketch are underlined with double lines. Other nominal objects that the Sketch Engine misses are marked with a dotted line.

- 6.a. In addition to encouraging kids to ask, think and do, parents need to be tolerant and appreciative to avoid killing a child's creative sense.
- b. Children are taught to love their parents, classmates, animals, nature in fact they are taught to love just about everything except to love China, their mother country.
- c. For example, the government deliberately chose not to teach Chinese history and culture, nor civics, in the schools.
- d. At the game there will be a lottery drawing for a motorcycle! And perhaps you'll catch a foul ball or a home run.

The sentences in (6) show that the current Sketch Engine tend to only identify the first object when there are multiple objects. The resultant distributional information thus obtained will be valid given a sufficiently large corpus. However, if the collocation patterns are fine-tuned to allow treatment of coordination, richer and more precise information can be extracted.

A regular expression collocation pattern also runs the risk of mis-classification. For instance, speech act verbs often allow subject to occur in post-verbal positions, and intransitive

verbs can often take temporal nouns in post-verbal positions too.

7. a. ...you can say goodbye to your competitive career.
 b. 'No,' said Scarlet, 'but then I don't notice much.'
8. a. Where did you sleep last night?
 b. ...it arrived Thursday morning.
 c. From Arty's room came the sound of an accordion.
9. 'I'll look forward to that.' 'So will I.'

Such non-canonical word orders are even more prevalent in Chinese. Chinese objects often occur in pre-verbal positions in various pre-posing constructions, such as topicalization.

10. 全穀麵包，吃了很健康。
quan.gu mian.bao, chi le hen jian.kang
 whole-grain bread, eat LE very healthy
 'Eating whole-grain bread is very healthy.'
- 11a. 有人嘗試要將這荷花分類，卻越分越累。
you ren chang.shi yao jiang zhe he.hua fen.lei
que yue fen yue lei
 someone try to JIANG the lotus classify, but more
 classify more tired
 'People have tried to decide what category the lotus belongs in, but have found the effort taxing.'
- b. 我一定要把老大除掉。
wo yi.ding yao ba lao.da chu.diao
 I must want BA the oldest (son) get rid of
 'I really want to get rid of the older son.'

When objects are pre-posed, they tend to stay closer to the verb than the subject. Adding object marking information, such as *ba* 把, *jiang* 將, *lian* 連 would help correctly identify collocating pre-posed objects. However, for those unmarked pre-posed structures, closeness to the verb may not provide sufficient information. Several rules will need to be implemented jointly.

The above example underlines a critical issue. That is, whether relative position alone is enough to identify positional information. The Sketch Engine is in essence a powerful tool extracting generalizations from annotated corpus data. We have shown that it can extract useful grammatical information with POS tag alone. If the corpus is tagged with richer annotation, the Sketch Engine should be able to extract even richer information.

The Sinica Corpus tagset adapts to the fact that Chinese has a freer word order than English by incorporating semantic information with the grammatical category. For instance, locational and temporal nouns, proper nouns, and common nouns each are assigned a different tag. Verbs are sub-categorized according to activity and transitivity. Such information is not available in the BNC tagset and hence not used in the original Sketch Engine design. We will enrich the collocation patterns with the annotated linguistic information from the Sinica Corpus tagset. In particular, we are converting ICG lexical subcategorization frames (Chen and Huang 1990) to Sketch Engine collocation patterns. These ICG frames, called Basic Patterns and Adjunct Patterns, have already been fully annotated lexically and tested on the Sinica Corpus. We expect their incorporation to improve Chinese Sketch Engine results markedly.

6. Conclusion

In this paper, we introduce a powerful tool for extraction of collocation information from large scale corpora. Our adaptation proved the cross-lingual robustness of the Sketch Engine. In particular, we show the robustness of the Sketch Engine by achieving better results through fine-tuning of the collocation patterns via integrating available grammatical knowledge.

References

- Chen, Keh-Jiann and Huang, Chu-Ren. 1990. Information-based Case Grammar. *Proceedings of the 13th COLING*. Helsinki, Finland. 2:54-59.
- Chen, Keh-Jiann, Chu-Ren Huang, Feng-Yi Chen, Chi-Ching Luo, Ming-Chung Chang, and Chao-Jan Chen. 2003. Sinica Treebank: Design Criteria, Representational Issues and Implementation. In Anne Abeill'e, (ed.): *Building and Using Parsed Corpora. Text, Speech and Language Technology*, 20:231-248. Dordrecht: Kluwer.
- CKIP (Chinese Knowledge Information Processing Group). 1995/1998. *The Content and Illustration of Academia Sinica Corpus. (Technical Report no 95-02/98-04)*. Taipei: Academia Sinica
- Huang, Chu-Ren, Feng-Ju Lo, Hui-Jun Hsiao, Chiu-Jung Lu, and Ching-chun Hsieh. 2001. From Language Archives to Digital Museums: Synergizing Linguistic Databases. Presented at the IRCS workshop on linguistic Databases. University of Pennsylvania.
- Huang, Chu-Ren, Keh-Jiann Chen, and Lili Chang. 1997. Segmentation Standard for Chinese Natural Language Processing. *Computational Linguistics and Chinese Language Processing*. 2(2):47-62.
- Kilgarriff, Adam and Tugwell, David. Sketching Words. 2002. In Marie-Hélène Corréard (ed.): *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*. 125-137. Euralex.
- Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychlý, Simon Smith, and David Tugwell. 2005. Chinese Word Sketches. ASIALEX 2005: Words in Asian Cultural Context. Singapore.
- Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz and David Tugwell. 2004. The Sketch Engine. *Proceedings of EURALEX*, Lorient, France. (<http://www.sketchengine.co.uk/>)
- Leech, Geoffrey. 1992. 100 million words of English: the British National Corpus (BNC). *Language Research* 28(1):1-13
- Lin, Dekang. 1998. An Information-Theoretic Definition of Similarity. *Proceedings of International Conference on Machine Learning*. Madison, Wisconsin. (<http://www.cs.umanitoba.ca/~lindek/publication.htm>)
- Tsai, Mei-Chih, Chu-Ren Huang, Keh-Jiann Chen, and Kathleen Ahrens. 1998. Towards a Representation of Verbal Semantics--An Approach Based on Near Synonyms. *Computational Linguistics and Chinese Language Processing*. 3(1): 61-74.
- Wu, Yiching and Liu, Mei-Chun. 2003. The Construction and Application of Mandarin Verbnet. *Proceedings of the Third International Conference of Internet Chinese Education*. 39-48. Taipei, Taiwan.
- Xia, Fei, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. *Proceedings of the second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece. (<http://www.cis.upenn.edu/~chinese/ctb.html>)
- Xue, Nianwen and Palmer, Martha. 2003. Annotating Propositions in the Penn Chinese Treebank. *Proceedings of the Second Sighan Workshop*. Sapporo, Japan. (<http://www.cis.upenn.edu/~xueniwen/>)
- Xue, Nianwen and Palmer, Martha. 2005. Automatic Semantic Role Labeling for Chinese Verbs. *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. Edinburgh, Scotland. (<http://www.cis.upenn.edu/~xueniwen/>)

Websites

- Sinica Corpus.
<http://www.sinica.edu.tw/SinicaCorpus/>
- British National Corpus (BNC).
<http://www.natcorp.ox.ac.uk/>
- Center for Chinese Linguistics, PKU.
<http://ccl.pku.edu.cn/#>
- Corpora And NLP (Natural Language Processing) for Digital Learning of English (CANDLE).
<http://candle.cs.nthu.edu.tw/candle/>
- FrameNet.
<http://www.icsi.berkeley.edu/~framenet/>
- Penn Chinese Treebank.
<http://www.cis.upenn.edu/~chinese/ctb.html>
- Proposition Bank.
<http://www.cis.upenn.edu/~ace/>
- Sinica Treebank.
<http://treebank.sinica.edu.tw/>
- Sketch Engine (English).
<http://www.sketchengine.co.uk/>
- Sketch Engine (Chinese).
<http://corpora.fi.muni.cz/chinese/>
- Sou Wen Jie Zi*-A Linguistic KnowledgeNet.
<http://words.sinica.edu.tw/>