A Continuum-Based Approach for Tightness Analysis of Chinese Semantic Units

Ying Xu[a], Christoph Ringlstetter[b], and Randy Goebel[a]

[a] Department of Computing Science, University of Alberta,
2-21 Athabasca Hall, Edmonton T6G 2E8, Canada
{yx2, goebel}@cs.ualberta.ca
[b] Center for Language and Information Processing (CIS), Ludwig Maximilians University
Oettingen Strasse 67, Munich 80538, Germany
kristoph@cis.uni-muenchen.de

Abstract. Chinese semantic units fall into a continuum of connection tightness, ranging from very tight, non-compositional expressions, tight compositional words, phrases, and then to loose more or less arbitrary combinations of words. We propose an approach to measure tightness connection within this continuum, based on document frequency of segmentation patterns in a reference corpus. A variety of corpora, including search engine snippets, search engine results derived from query logs, as well as standard corpora have been investigated. Our tightness ranking on 300 phrases is quite close to their manual ranking, and non-compositional compound extraction can achieve a precision as high as 94.3% on the top 1,000 4-grams extracted from the Chinese Gigaword corpus.

Keywords: Compounds in Mandarin, Collocation, Compositionality.

1 Introduction

Many people are working on acquisition of multi-gram semantic units, although the terminology varies. “Gram” here means “sociological word,” which is the familiar “word” in English, and the “character” in Chinese (Packard, 2000). Whether the goal is collocation extraction (Lin, 1998), multiword expression extraction (Sag *et al.*, 2002), or Chinese word extraction (Feng *et al.*, 2004; Xu and Lu, 2006), they all try to extract multi-gram semantic units for which the meaning as a whole can not be predicted from the meaning of the gram units, and called “limited compositional” or “non-compositional.” Multi-gram extraction identifies strings like “kick the bucket,” “at gunpoint,” or “make out” in English, and strings like “花生” (peanut), “月下老人” (match maker), or “乌鲁木齐” (Urumchi, name of a city) in Chinese.

Multi-gram extraction is important for many natural language processing (NLP) applications. For example, it can be used for lexicon acquisition from corpora, extracting new words like “正龙射虎” (metaphor for cheating), which appeared in the Internet after 2008; it can be used for a word-based indexing of an information retrieval (IR) system; furthermore, it can be beneficial for word-based machine translation (MT). The impact of Chinese word extraction (or segmentation) on the last two tasks has been intensively analyzed (Nie *et al.*, 2000; Foo and Li, 2004; Peng *et al.*, 2002; Chang *et al.*, 2008). The results suggest the relationship is not monotonic, better segmentation does not always yield better MT or IR performance.

Our hypothesis for this phenomenon is that independent Chinese semantic units (also referred to as “Chinese strings” in the following) as observed in a text do not fall cleanly into the binary classes of compositional or non-compositional, but into a continuum of tightness and looseness, where tightness is considered as a degree of compositionality. Intuitively, this continuum also exits in naturally segmented languages such as English (Halpern, 2000). This tightness characteristic of strings determines their linguistic nature as well as their preferred treatment in different NLP applications, e.g., for two consecutive nouns, whether to index two nouns or one nominal compound in IR, or to translate them as a unit or separately. For different NLP applications, the threshold for how tight a Chinese string need to be so that we keep it as a word will be different, but binary classification of semantic units is not enough.

On this tightness continuum, at one extreme are non-compositional semantic units, such as idioms, non-compositional compounds, and transliterated names; at the other end are purely consecutive words which means there is no dependency relation between those words, with compositional compounds and phrases in between. Figure 1 shows some examples of English and Chinese multi-gram semantic units along this tightness continuum, where the left end is tightest and the right end is loosest. For English, "going Dutch" is a non-compositional idiomatic expression as its meaning has nothing to do with combination of the literal meanings of "going" and "Dutch"; the same holds for "milky way," a non-compositional compound; "machine learning" is a compositional compound but a tight one as compared to "plum pie" which is significantly looser; "last year" is a common sense phrase with "last" as a modifier of "year"; "CPU can't" is a phrase in a text with an arbitrary nominal CPU preceding the very general modal "can." For Chinese, "月下老人" (match maker) is a non-compositional idiomatic expression since its meaning has nothing to do with combination of the literal meaning of "月下" (under the moon) and "老人" (old people); "乌鲁木齐" (Urumchi) is a non-compositional transliterated proper noun; "机器学习" (machine learning) is a compositional compound; "正当收入" (legitimate income) is a phrase; and "上海哪有" (Shanghai where) are two consecutive words.

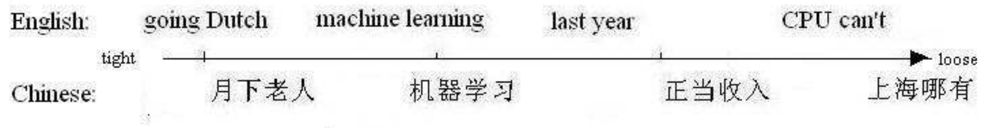

Figure 1: continuum of tightness.

In our work, we exploit corpus data and propose a method to locate a Chinese character string in the continuum of tightness and looseness. The input of our approach are document frequencies of segmentation patterns for strings in corpora, i.e. number of documents that contain a specific segmentation pattern. A pattern is a potential segmentation, which here means that a character string of length n has 2^{n-1} different patterns. For example, "|机器学习|", "|机|器学习|" and "|机器|学习|" are possible segmentation candidates for "机器学习" (machine learning). Note that every pattern contains all the characters of the string. The intuition of using document frequency is that a document that contains all the units of a string provides a stronger basis for the semantics of that string than a document that, for example, contains only one unit.

We confirm the value of our approach with two experiments. First, we use our method to rank 300 Chinese strings according to their tightness and compare that result with a manually created gold standard ranking. The evaluation shows the automatic ranking is comparable to the manual ranking. Second, we extract non-compositional semantic units from the Chinese Gigaword corpus and compare the result with a dictionary. The precision is promising, which further supports the value of our tightness continuum measure.

Our paper is organized as follows. Section 2 introduces the related work on tightness measures and Chinese word extraction. Section 3 presents our approach to measure the tightness of strings built from consecutive Chinese characters. In Section 4, we present the evaluation procedure and results. A brief conclusion summarizes our findings and anticipates future work.

2 Related Work

There have been a number of methods proposed for extracting multi-gram semantic units, and for measuring the tightness of multiword expressions in linguistic studies (Bannard *et al.*, 2003; McCarthy *et al.*, 2003; Kim and Baldwin, 2007). Much of this work has proposed measures for the tightness of multiword expressions in English, while there are very few such Chinese word extraction methods.

Most of these collocation extraction methods use or are related to pointwise mutual information (MI), which is one of the most popular ways to extract collocations or compounds (Manning and Schutze, 1999). The standard approach is to conceive the random variables of MI as lexical items, and approximate the probabilities of those random variables by counting lexical items in a corpora. So one can apply the concept of MI between lexical items x and y as follows:

$$MI(x,y) = \log \frac{P(xy)}{P(x)P(y)} \tag{1}$$

where $P(x)$ is the probability of x in a corpus, $P(xy)$ is the probability that x and y are consecutive within that corpus. This measure has been used for multi-gram extractions in both English and Chinese. As explained below, one difference in our tightness measure is in how we use counting in a corpus to approximate the probabilities that define our measure, as compared to lexical versions of MI. In our case, the denominator is calculated by counting non-adjacent occurrences of both x and y occurring within one document (see details below). This is a better way to catch semantic meaning, for example, if “machine” occurs in one document but not “learning”, then that “machine” can be a car engine, a copy machine, other than a computer. Generally speaking, this is a way of word sense disambiguation which MI ignores.

Lin (1998) presented a method for non-compositional English phrase extraction based on the hypothesis that when a phrase is non-compositional, its mutual information differs significantly from the mutual information of phrases obtained by substituting one of the words in the phrase with a similar word. He compared the results with two manually compiled English dictionaries: in one precision and recall is 15.7%, 13.7%; in the other, precision is 39.4%, and recall is 20.9%. This shows that even lexicographers can disagree about which phrases are non-compositional.

McCarthy *et al.* (2003) investigated various statistical measures of compositionality of candidate multiword verbs, specifically English phrasal verbs identified automatically using a robust parser. Note that their work depends on the result of automatic part of speech (POS) tagging and a synonym list, but our method takes raw corpus data as input directly. The best result they got is a correlation of 0.49 with human annotators. While they ranked their test phrase set on a 10 rank scale, we ranked them on a 3 rank scale, since it is more difficult even for human annotators to rank a phrase when the scale is more fine-grained, and it is not necessary for NLP applications such as IR and MT.

There is a lot of research which employs statistical methods to extract Chinese words or to segment Chinese. In (Feng *et al.*, 2004) a method based on statistical data called “context variety” was employed to extract candidates. The idea is to consider the variance of characters appearing on the right and left sides of a target character. Strings with high variety are extracted, as such strings appear in enough different environments to have the potential to be meaningful. They measured their extraction word list by comparing with a Chinese dictionary and calculated the precision. But this method, as with many other Chinese word extraction methods, did not consider whether word units are compositional or not. For example, they extracted units such as “假冒伪劣商品” (fake and bad merchandise), which is not clearly a compositional word or a phrase.

(Xu and Lu, 2006) is one of the few studies, which classify Chinese collocations according to their tightness. In this case tightness is distributed over 4 classes: idiomatic collocations such as “缘木求鱼” (to climb a tree to catch a fish, meaning a fruitless effort), fixed collocations such as “外交豁免权” (diplomatic immunity) in which two components can not be substituted by other words to carry the same meaning, strong collocations such as “缔结同盟” (form alliance) with limited modifiability, and loose collocations such as “合法收入” (lawful income) of which the replacement of components is not arbitrary. The input corpus, from which they extracted collocation candidates, is segmented and POS tagged. They evaluated the extracted collocation precision according to a manually extracted set. One difference between our method and theirs is that our method locates Chinese semantic units in a continuous spectrum, while they classify them into 4 classes. Our method can be more dynamic to meet different application needs, as generally it is difficult to separate between fixed collocations and strong collocations, between strong collocations or loose collocations. For example, “machine learning” may be a compound to some people, but may be a phrase to others. Another important difference is that their method is based on a large segmented and POS-tagged corpus, while our method is based on a large raw corpus. If a string of compositional phrases such as “假冒伪劣商品” (fake and bad merchandise) is wrongly segmented as a unit, then their method can not classify them correctly.

3 Computing the tightness continuum

We locate a Chinese string within the tightness continuum by a new measure, of which the input is the probability distribution of the string’s patterns, i.e., potential segmentation candidates; the output is a continual tightness value: the greater the value, the tighter the string.

3.1 Pattern frequency

Here we introduce how we get the input of the measure, i.e. pattern frequencies. As mentioned before, for a string of length n exit 2^{n-1} potential segmentation candidates. In case of a 4-gram “ABCD”, there are 8 candidates: Pt(ABCD), Pt(A|BCD), Pt(AB|CD), Pt(ABC|D), Pt(A|B|CD), Pt(A|BC|D), Pt(AB|C|D), and Pt(A|B|C|D), where ‘|’ is used as a segmentation delimiter. Each candidate is called a potential *pattern*. Note that typically only a subset of the patterns is linguistically valid. In the following, we give a detailed description of patterns for a 4-gram “ABCD”. First we introduce the component patterns for its 8 segmentation patterns. The used regular expression language is in Java notation, and ≜ means “mark as.”

- “[ˆA]BCD” ≜ Pt(BCD): a string with “BCD” without character “A” in front of “BCD”. Take 4-gram “机器学习” (machine learning) as an example, string “关于机器学习”, “学习写作” do not match with this pattern, as “机” is in front of “器学习” for the first one and “器” is missing for the second one; while “计算器学习” does match with this pattern.
- “ABC[ˆD]” ≜ Pt(ABC): a string with “ABC” without character “D” following “ABC”. Take 4-gram “机器学习” as an example again, string “关于机器学习”, “机器人” do not match with this pattern, while “机器学做菜” does.
- “AB[ˆC]” ≜ Pt(AB): similar to Pt(ABC).
- “[ˆB]CD” ≜ Pt(CD): similar to Pt(BCD).
- “A[ˆB]” ≜ Pt(A): similar to Pt(ABC).
- “[ˆA]B[ˆC]” ≜ Pt(B): a string with “B” without “A” in front of “B” and without “C” following “B”. Take “机器学习” as an example again, “机器”, “计算器学习” do not match with this pattern, while “计算器计算” do.
- “[ˆB]C[ˆD]” ≜ Pt(C): is similar to Pt(B).
- “[ˆC]D” ≜ Pt(D): is similar to Pt(BCD).

Having introduced the components of the 4grams’ 8 patterns, in what follows we describe how the 8 patterns are counted.

- Pt(ABCD): if the whole string appears in one document, then we say the document is evidence for this pattern and the frequency count of Pt(ABCD) is incremented by 1.
- Pt(A|BCD): if Pt(BCD) and Pt(A) are inside a document, then we say the document is evidence for this pattern and the count of Pt(A|BCD) is incremented by 1. Take 4-gram “机器学习” as an example again, string “关于机器学习”, “计算器学习” do not match with this pattern, as “机” is in front of “器学习” for the first one and Pt(A) is missing for the second one; while “机动车计算器学习” does match with this pattern.
- Pt(AB|CD): if Pt(AB) and Pt(CD) are inside a document, then we say the document is evidence for this pattern and the count of Pt(AB|CD) is incremented by 1.
- Pt(ABC|D): if Pt(ABC) and Pt(D) are inside a document, then we say the document is evidence for this pattern and the count of Pt(ABC|D) is incremented by 1..
- Pt(A|B|CD): if Pt(CD) is in a document and the document contains Pt(A) and Pt(B), then we say the document is evidence for this pattern and the count of Pt(A|B|CD) is incremented by 1. Take 4-gram “机器学习” as an example again, string “机动车学习”, “机器人学习” do not match with this pattern, as Pt(B) is missing for the first one and both Pt(A) and Pt(B) are missing for the second one; while “机动车计算器的学习” does match with this pattern.
- Pt(A|BC|D): if a document contains Pt(BC), Pt(A), and Pt(D), then we say the document is evidence for this pattern and the count of Pt(A|BC|D) is incremented by 1.
- Pt(AB|C|D): similar to Pt(A|B|CD).
- Pt(A|B|C|D): if a document contains Pt(A), Pt(B), Pt(C), and Pt(D) then we say the document is evidence for this pattern and the count of Pt(A|B|C|D) is incremented by 1. Take 4-gram ”机器学习” as an example again, string “计算器机动车学自习” matches this pattern.

Whenever one of the 8 segmentation patterns occurs in a document, that document is evidence for the pattern, and the frequency count of the pattern is incremented by 1. One document can be evidence for several patterns. For example, for 4-gram ”机器学习”, string “机器学习领域训练机器自主学习” is evidence of Pt(ABCD) and Pt(AB|CD).

3.2 Tightness Measure

We assume a string is tight with respect to a chosen corpus if, when all component characters of a string appear in a document, they are more likely to appear in one consecutive form, i.e. in the form of the string. So the more frequent the whole string pattern is compared to other patterns which separate the component characters, the tighter the string is. Consider the 4-gram “ABCD” again, the more frequent the Pt(ABCD) is compared to other 7 patterns, the tighter “ABCD” is. In contrast to MI, where frequencies of parts x and y are typically based on the whole corpus, our method only considers those documents where both x and y appear. We do this to avoid insignificant counts of documents with only x or y that do not relate to the semantics of the whole string. Besides, while MI considers term frequency, our method only considers document frequency, because it is difficult to tell whether the appearance of “AB” is evidence of Pt(AB|CD) or Pt(AB|C|D).

We consider only patterns that segment a string into two parts, which means we do not consider Pt(AB|C|D), Pt(A|B|CD), or Pt(A|B|C|D), etc. One reason is because the greater order a gram is, i.e. the longer a gram is, the better it can hold specific semantic intention. A document with “医生” (doctor) and “护士” (nurse) will have a greater chance to be related to “医生护士” than a document with “医”, “生”, “护”, and “士”. Another motivation is that the patterns in which a 4-gram separates into two parts are also observations about the other three part or four part segmentation candidates. For example, if a 4-gram can be segmented into |AB|C|D|, then observations Pt(ABC|D) and Pt(AB|CD) are also possible; if a 4-gram can be segmented into |A|B|C|D|, then all the 8 pattern observations are possible. Take “我很想你” (I miss you very much) as an example, it can be segmented into “|我|很|想|你|”, so we expect Pt(A|B|C|D), Pt(AB|C|D), Pt(A|B|CD), etc. will occur. For “机器学习”, the semantically reasonable segmentation is “|机器|学习|”, Pt(A|B|C|D) will be rare as compared to Pt(AB|CD). We assume that that the more parts a 4-gram can be separated into, the looser it is.

Among patterns that segment a string into two parts, we assume the most frequent one is the most semantically reasonable one. For “机器学习”, we expect Pt(机器|学习) will be more frequent than Pt(机|器学习) or Pt(机器学|习).

With these observations, we propose the following tightness measure,

$$ratio = \begin{cases} \frac{\sharp Pt(\text{whole string})}{\max(\sharp Pt(\text{patterns segmenting string into two parts}))+\frac{1}{N}} & \text{if } \sharp Pt(\text{whole string}) > \sigma \\ \text{undef} & \text{otherwise} \end{cases} \tag{2}$$

where $\sharp$ means frequency, σ is a threshold to exclude rare patterns, which is set as 50 in the following experiment, and N is a smoothing factor which is set as the number of documents. So for 4-grams, the function will be,

$$ratio = \begin{cases} \frac{\sharp Pt(ABCD)}{\max(\sharp Pt(A|BCD),\sharp Pt(AB|CD),\sharp Pt(ABC|D))+\frac{1}{N}} & \text{if } \sharp Pt(ABCD) > \sigma \\ \text{undef} & \text{otherwise} \end{cases} \tag{3}$$

This tightness measure can be used to compare tightness between strings. Moreover, we can set a threshold of the value, and assume grams with tightness above the threshold as non-compositional when we extract Chinese semantic units.

4 Experiments

Our hypothesis is that there is a continuum of tightness for Chinese strings, and it can be modeled by the measure we proposed. To prove that our measure does catch the tightness of Chinese strings, we conduct two experiments. First, we use our method to rank 300 4-gram Chinese strings, which include non-compositional words, compositional words, and phrases, according to their tightness. We then compare the result with a manually created gold standard ranking. In the second experiment, we rank all the 4-grams in the Chinese Gigaword corpus according to their tightness, and assume the top 3,000 are non-compositional semantic units, such as idioms and transliterated names. We then compare these 3,000 grams with a dictionary which we assume is a non-compositional compound list. Note that our method is not only limited to 4-grams; we choose 4-grams as an example because 4-gram compounds are more prominent than others, just as bi-gram words which are prominent in Chinese.

4.1 Rank Similarity

In the following experiment, we compare 300 4-grams’ ranking according to their tightness by using our measure and MI with manual ranking. The 300 4-grams appear in Sogou query logs of March 2007 and are tagged as noun phrases in the Chinese Treebank (Xia *et al.*, 2000). The use of the Treebank ensures that the 300 grams are complete meaningful units. In order to analyze the influence of different corpora, we employ five web-based corpora and one standard corpus, the Chinese Gigaword.

- 4 sets of snippets from 4 Chinese search engines, Baidu, Sogou, Google, and Yahoo!. We query the search engines for the 300 4-grams and recorded about 500 snippets for each query. For example, we sent “可口可乐” (Coca-Cola) to Baidu, record the first 500 snippets, and calculated the frequency of Pt(可口可乐), Pt(可|口可乐), etc., based on these 500 snippets.

- Web pages clicked in the Sogou query logs where the 300 phrases matches a user query or part of a user query. (cf. Table 1 that shows a sample piece of the Sogou query logs. The respective documents have been downloaded for the experiment.) The first record in the table is for query [南粤双色球开奖结果] posed at 00:00:00 by user 34217485189702995. URL “www.0769888.com/qsc0769/849934712.html” ranks third by the Sogou search engine for that query and is the first URL the user clicked for that query. If one of the 300 phrases is “网易聊天” (Wangyi Chatting), then web page “chat.163.com/” will be considered as a support document for “网易聊天” because of the third record of the query log in Table 1.

Table 1: Sample piece of Sogou query logs.

00:00:00 34217485189702995 [南粤双色球开奖结果] 3 1 www.0769888.com/qsc0769/849934712.html
...
00:00:04 34217485189702995 [南粤双色球开奖结果] 3 2 www.0769888.com/qsc0769/849934712.html
00:00:04 34062155775183716 [网易聊天室] 1 1 chat.163.com/
00:00:04 04324790273288531 [西安婚纱道具] 8 1 dzh.mop.com/topic/readSub_6280165_0_0.html
...
00:00:12 04324790273288531 [西安婚纱道具] 9 2 www.029apple.com/newforum/hAnnounceShow.asp?HFA_ID=30862&nCurpage=1
...
00:01:01 04324790273288531 [西安婚纱道具] 18 3 vip.wedchina.com/bbs/dispbbs.asp?boardID=41&ID=138380&page=1

- The Chinese Gigaword Corpus. In order to get pattern distributions of 4-grams from the Chinese Gigaword corpus, we need to extract documents that contain substrings of 4-grams. So we build inverted indices for unigrams, bigrams, trigrams, and 4-grams in the Gigaword corpus, using an open source Lucene package (Hatcher and Gospodnetic, 2004). [1]

For the web-based corpora we filter out 4-grams from our test set whose sum of the frequencies of all possible segmentation patterns is less than 50; for the Gigaword corpus we filter out all 4-grams where the frequency of the consecutive pattern Pt(ABCD) is less than 50. The number of 4-grams left for every corpus is in Table 2. We calculate the tightness value of 4-grams from these corpora, and sort them in descending order based on this value. So rank 1 will be the tightest.

Table 2: number of 4-grams with ratio defined.

corpus	Baidu	Google	Sogou	Yahoo!	Web pages	Gigawd
|4-grams|	297	300	266	295	230	283

To find the difference between our method and pointwise mutual information, we also rank the 4-grams by point mutual information according to the Chinese Gigaword corpus. To compute a 4-gram’s mutual information, we segment it into two parts according to the patterns’ frequencies. For example, for a gram “ABCD”, if $\max(\sharp Pt(A|BCD), \sharp Pt(AB|CD), \sharp Pt(ABC|D)) = \sharp Pt(A|BCD)$, then part1 = “A”, part2 = “BCD”. So the mutual information of a 4-gram is,

[1] The code is available at http://www.cs.ualberta.ca/˜yx2/pattern.zip.

$$\log \frac{\sharp' Pt(ABCD) * N}{\sharp' Pt(part1) * \sharp' Pt(part2)} \tag{4}$$

where $\sharp' Pt(i)$ means total term frequency, not just document frequency, and N is number of words in the Chinese Gigaword corpus (approximately 10^8).

To create a gold standard ranking, a human annotator ranked the 300 phrases of the test set on a 3 rank scale: rank 1 means very tight, for example, idioms or transliterated proper nouns, “澳大利亚” (Australia), “花花公子” (playboy); 2 means tight, such as compositional compounds, “人民银行” (people bank), “哈尔滨市” (Harbin city); and rank 3 denotes general phrases.

We use Kendall’s τ to compare two ranks (Kendall, 1955):

$$\tau(r_a, r_b) = \frac{P - Q}{P + Q} \tag{5}$$

where P is number of equal values between two ranks r_a and r_b, and Q is number of different values between two ranks. For comparison between automatic rankings, $P + Q = \binom{n}{2}$, where n is the size of intersection between two ranking domains. For example, for rankings based on Baidu and Google, there are 297 grams in the intersection. For comparison between an automatic ranking and the manual ranking, $P + Q = \frac{n_1 * n_2}{2} + \frac{n_1 * n_3}{2} + \frac{n_2 * n_3}{2}$, where n_i is number of 4-grams in rank i set. We do not distinguish grams falling in the same rank in this case as it is difficult to decide which is more tight, e.g., an idiom “一枝独秀” (outshine others), or an idiom “白手起家” (start from scratch).

Table 3 shows the similarities of aforementioned tightness ranks against different corpora. “Baidu” means the ranking using our tightness measure against Baidu search engine snippets. “Gigawd_Ratio” means the ranking using our tightness measure against the Chinese Gigaword corpus. “Gigawd_MI” means the ranking using MI measure against the Chinese Gigaword corpus.

Table 3: Rank similarities of measurements.

	Baidu	Google	Sogou	Yahoo!	Web pages	Gigawd_Ratio	Gigawd_MI
Baidu	\	0.71	0.73	0.73	0.70	0.45	0.39
Google	\	\	0.74	0.74	0.72	0.45	0.39
Sogou	\	\	\	0.76	0.72	0.48	0.41
Yahoo!	\	\	\	\	0.74	0.48	0.41
Web pages	\	\	\	\	\	0.50	0.43
Gigawd_Ratio	\	\	\	\	\	\	0.73
Gigawd_MI	\	\	\	\	\	\	\
Manual rank	0.69	0.66	0.65	0.72	0.66	0.58	0.42

The result shows more similarity between the automatic ranking using our approach and the manual ranking as compared to the ranking using MI, which provides evidence that our method measures the tightness of Chinese strings in a reasonable way. MI ranks collocations such as “妇幼保健” (maternity and child care) high, i.e. very tight, even higher than the transliteration “马来西亚” (Malaysia), while our method ranks the former lower than the latter. The Chinese Gigaword corpus obtains the lowest similarity, which we believe is simply because it is small relative to the size of the documents indexed by the search engines, so is consistent with the general intuition about the emergent accuracy of simple statistics applied to large data sets.

4.2 Non-compositional units extraction

In the second experiment, we extract non-compositional 4grams from the Chinese Gigaword corpus according to our tightness measure. The gold standard is a dictionary combined with a human judge for those 4-grams that are not in the dictionary. This means we assume a candidate that is in the dictionary to be non-compositional which holds for most cases but not for all. For example, we found the compositional expression “道德品质” (moral trait) in the dictionary. Nevertheless the use of the dictionary produces more objectivity as compared to the approach if we had all 4-grams ad hoc judged by only one human.

We rank all 4-grams in the Chinese Gigaword corpus according to the tightness measure and analyze the first 3,000 4-grams, which we assume as non-compositional Chinese semantic units, out of a total 830,809 4-grams. First we try to find these 3,000 4-grams in the “Modern Mandarin word dictionary.” If a 4-gram is not in the dictionary (it is neither a lexical item in the dictionary nor part of a lexical item), we query it on the Baidu search engine to check manually if it is a proper noun, e.g., person names, and location names. If a 4-gram is in the dictionary or a non-compositional proper noun, for example proper nouns like “上海市”(Shanghai city) are compositional, but nouns like “上海” (Shanghai) are, then it is correctly extracted. The precision for the first 1000 4-grams is 94.3%; the precision for the first 2000 is 89.5%; and the precision for the first 3000 is 81.1%.

To compare our method with MI, we rank all 4-grams according to the standard lexicalized version of MI. The precision calculated according to the proposed evaluation for the first 1000 4-grams is 66.3%. When we analyze the 4-grams manually, we find the MI measure extracts more loose collocations and fixed compositional expressions. Examples are expressions like “虚报浮夸” (make a false report, exaggerate), “公道正派” (just, honest), “叔叔阿姨” (uncle, aunt) which are ranked high by MI, but low by our measure.

5 Conclusion

Our hypothesis is that Chinese strings as observed in a text fall into a continuum of tightness and looseness. We proposed a tightness measure to locate Chinese semantic units within this continuum based on the statistical distribution of their component characters in a variety of corpora. Tightness ranks of phrases by our measure based on different corpora, including search engine snippets, web pages, and the Chinese Gigaword corpus, show high similarity with human rank. A second experiment related to the extraction of non-compositional expressions showed promising results when we evaluated the precision of our method against a dictionary as compared to MI.

Besides its value for linguistics, we believe our approach can benefit applications such as machine translation and information retrieval. If a string is non-compositional, the IR system should treat it as a word; if a string is loose, it should be segmented. We can analyze IR performance employing segmentations based on different thresholds of how tight a string needs to be, to be considered as non-compositional. For example, there is no doubt that “月下老人” (match maker) is a single word and “上海哪有” (Shanghai where) are two words, but what about “机器学习” (machine learning), should it be segmented for IR or kept as a word? Such experiments will help understanding the effects of Chinese word segmentation upon NLP tasks.

References

Bannard, C., T. Baldwin and A. Lascarides. 2003. A Statistical Approach to the Semantics of Verb-Particles. *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions*, pp.65-72.

Chang, P., M. Galley and C.D. Manning. 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. *Proceedings of the Third Workshop on Machine Translation.*

Feng, H., K. Chen, X. Deng and W. Zheng. 2004. Access Variety Criteria for Chinese Word Extraction. *Computational Linguistics*, 30(1).

Foo, S. and H. Li. 2004. Chinese word segmentation and its effect on information retrieval. *Information Processing and Management: an International Journal*, 40(1).

Halpern, J. 2000. Is English Segmentation Trivial? *Technical report, CJK Dictionary Institute*.

Hatcher, E. and O. Gospodnetic 2004. *Lucene in Action*. Manning Publications Co.

Kendall, M. 1955. *Rank Correlation Methods*. Hafner.

Kim, S.N. and T. Baldwin. 2007. Detecting Compositionality of English Verb-Particle Constructions using Semantic Similarity. *Proc of the 10th Conference of the Pacific Association for Computational Linguistics*.

Lin, D. 1998. Automatic Identification of Non-compositional Phrases. *In Proceedings of the 37th Annual Meeting of the ACL,* 317-24, College Park, USA.

Manning, C.D. and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

McCarthy, D., B. Keller and J. Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. *Proc. Of the ACL-SIGLEDX Workshop on Multiword Expressions*.

Nie, J.Y., J. Gao, J. Zhang and M. Zhou. 2000. On the use of words and N-grams for Chinese information retrieval. *Fifth International Workshop on Information Retrieval with Asian Languages*.Hong Kong.

Packard, J.L. 2000. *Morphology of Chinese: A Linguistic and Cagnitive Approach*. Cambridge University Press.

Peng, F., X. Huang, D. Schuurmans, and N. Cercone. 2002. Investigating the Relationship between Word Segmentation Performance and Retrieval Performance in Chinese IR. *Retrieval Performance in Chinese IR, Coling2002*.

Sag, I.A., T. Baldwin, F. Bond, A. Copestake and D. Flickinger. 2002. Mutliword Expression: A Pain in the Neck for NLP. *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*.

Xia, F., M. Palmer, N. Xue, M.E. Okurowski, J. Kovarik, F.D. Chiou, S. Huang, T. Kroch, and M. Marcus. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. *Proc. of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Xu, R. and Q. Lu. 2006. A Multi-stage Chinese Collocation Extraction System. *Lecture Notes in Computer Science, Vol. 3930*, pp.740-749. Springer.