Targeting Chinese Nominal Compounds in Corpora

Weiruo Qu, Christoph Ringlstetter, Randy Goebel

Center for Information and Language Processing (CIS), Alberta Ingenuity Center for Machine Learning (AICML)
University of Munich, University of Alberta
quweiruo@cis.uni-muenchen.de, kristof@cs.ualberta.ca, goebel@cs.ualberta.ca

Abstract

For compounding languages, a great part of the topical semantics is conveyed via nominal compounds. Various applications of natural language processing can profit from explicit access to these compounds, provided by a lexicon. The best way to acquire such a resource is to harvest corpora that represent the domain in question. For Chinese, a significant difficulty arises because the text comes as a string of characters, segmented only by sentence boundaries. Extraction algorithms that solely rely on context variety do not perform precisely enough. We propose a pipeline of filters that starts from a candidate set established by accessor variety and then employs several methods to improve precision.

1. Introduction

For compounding languages, a great part of the topical semantics is conveyed via nominal compounds. Persistent arising semantic concepts are largely represented by new combinations of preexisting lexemes. Concerning the interoperability of language resources, this phenomenon can be useful because translations of the compounds of a text strongly support multilingual access for applications like document browsing or retrieval. The importance of compounds for *Search* can be illustrated by the following example. With the unidentified compound *French Revolution*, a search for *French* $\cap$ *revolution* led to a huge number of unwanted results referring to the promised neo-liberal reforms at the eve of the recent presidential election in France, instead of the historic event (Hansel, June 3rd 2007).

Unfortunately, even mono-lingual natural language processing (NLP) resources, as for example simple lexica, often neglect compounds. For less common areas of interest, in most cases, all specialized compounds are missing. This is especially true for Chinese that, as a prototype of an isolating language, nearly without exception, uses compounding to represent new concepts. For Chinese NLP, unrecognized compounds can cause a complete disfiguration of the intended semantics. So for successful operation it is essential that an NLP system has an electronic lexicon of compounds at its disposal.

Any construction of such a specific lexical resource has to begin with a list of compound candidates. Corpus driven methods can be used to acquire candidates that include the important concepts of a specific area and simultaneously minimize noise. As for the collection of a domain specific corpus via the API of search engines, Web based methods are well established (Kilgarriff and Grefenstette, 2003; Ringlstetter et al., 2007). The acquisition of compound candidates from such a corpus has to be automatized as far as possible. Complicated even for "segmentation free" languages as English (Halpern, 2000), any such approach for Chinese has to deal with the specific difficulties that evolve from the hidden word boundaries that characterize Hanzi text. The aim of this paper is to provide a collection pipeline that uses an arbitrary corpus as input and isolates a list of nominal compound candidates.

Our contributions are the following.

1. Evaluation of context variety to detect compound candidates;
2. Filtering remaining numeric compounds, Chinese and transliterated names;
3. Separate phrases and other than nominal compounds; and
4. Distinguish nominal compounds and pseudo nominal compounds using an SVM-classifier.

First, we provide a working definition of the compound phenomenon and its different gestalts in Mandarin Chinese. In Section 3. we present a basic algorithm and a cascade of filters that improve over mere context variety to detect compounds. Section 4. reports on evaluation experiments for the proposed methods. The conclusion summarizes the results.

2. Compounds in Mandarin Chinese

Both from a theoretical point of view as well as from an engineering point of view, the definition of a compound and the level of compositionality are constitutive for the construction of a compound dictionary. The morphological phenomenon of compounding is widely discussed in the literature of natural language processing (Halpern, 2000; Packard, 2000; Sag et al., 2002). Work has been done on compounding for a variety of languages, in order to develop their systematic description and the type of information that should be attached to form meaningful lexicon entries (Viegas et al., 1998; Guenthner and Blanco, 2004; Violeta et al., 2004; Melcuk, 1995).

2.1. A working definition of a Chinese Nominal Compound

What constitutes a compound (复合词) in Mandarin Chinese has been a matter of ongoing discussion. In classic texts, Chinese has been characterized as a monosyllabic language; a compound has often been described as a combination of two or more bound morphemes (Starosta et al., 1998). We agree with (Li and Thompson, 1989) that

this characterization of Chinese is no longer valid. Compounding is thus to be seen as a process that involves single morphemes or combinations of morphemes that constitute free morpho-syntactical units, i.e., words. These words are combined to form a new syntactically simple unit; a *nominal compound* is a resulting word that falls into the part of speech category, noun.[1] In (Zhan, 2000), a similar definition is given, proposing that a compound is composed of two or more roots(词根) where a root consists of one or more Chinese characters. Examples for roots are: 人(human), 雷达(radar), and 巧克力(chocolate).

A recent phenomenon, that also should be considered among compounds, concerns conglomerates of Chinese words with transliterated (e.g., 因特网 internet, 迷你裙 miniskirt) or Latin script parts (e.g., RAF-恐怖份子). With regards to the combinatorical upper bound, we found that by far most compounds are of a length less than six characters. We follow this restriction in our experiments. To summarize, we consider a nominal compound as a polysyllabic noun that can be segmented into two or more morphemes, either free morphemes or loan morphemes.[2]

2.2. Compositionality of compounding in Mandarin Chinese

In Mandarin Chinese compound word formation, we can distinguish different levels of relatedness between the word semantics of the whole compound and its parts. At the extremes we have, ideally, compositional and non-compositional compounds. For the latter, we get no hint of the meaning of the compound by available information about the meaning of its constituting parts. Examples are given in Table 1.

From a lexicalist perspective, research on features that determine the degree of compositionality seems to be promising; but in practice, compounds should be treated as non-compositional (Halpern, 2006). Consequently, as many as possible should be kept in a lexicon that is based on real world data for the domain in question.

3. A hybrid method for the detection of Chinese nominal compounds in corpora

In this section, we integrate a variety of approaches that isolate nominal compounds in document corpora to create a hybrid method with significantly improved performance, especially for sparse data. If, in the following, the expression *character string* is used, we refer to an arbitrary sequence of Chinese characters, not interrupted by a blank or a sentence boundary.

3.1. Basic context variety step

In (Feng et al., 2004) a method was introduced that uses the degree of context variety to statistically determine a combination of characters as a compound and then filters against a list of adjacent characters to exclude phrases that combine a word and a grammatical marker. The context variety for each character string is measured by the following attribute list:

S: = Begin of sentences
E: = End of sentences
L: = Predecessors of the string
R: = Successors of the string
LAV: = L + S, Number of left contexts
RAV: = R + E, Number of right contexts
RV : = |LAV - RAV|, Parameter for balance
AV : = min {LAV; RAV}, Accessor variety

Given a corpus of text, an exhaustive algorithm generates a database that stores all possible segmentation variants. To establish candidacy as a compound, the AV-value of a character string has to stay above a threshold, whereas the RV-value, measuring bias of the contexts to one side, has to stay below a threshold. That is, an approved character string has to occur within a minimum number of different contexts and has to be reasonably balanced between the left and the right hand side. The main contribution of Feng et al. was then to additionally filter phrasal combinations of a regular word and predecessing (们), succeeding (小) or delimiting grammatical characters (的). For example, the sequence 的人们has a high AV-value because it can be predecessed by a great variety of adjectives. The algorithm then filters sequences that contain an adjacent character, if the unit without those characters was already collected. Because of the filtering with a list of adjacent characters, the method is superior as compared to entirely statistical methods. However, major precision problems exist, concerning names, numeric compounds, and phrases, with a negative correlation of available data and noise (cf. Section 4.).

3.2. Refinement postsegmentation step

A first improvement of the adjacent character method in terms of precision is based on a rule augmented segmentation algorithm; since compounds are composed of free morphemes they have to be segmentable. One of the major problems of word segmentation (and also of the isolation of compound words) is the recognition of personal names

Hanzi	Pinyin	Literal	English
花生	hua-sheng	flower-born	peanut
肥皂	fei-zao	fat-black	soap
薪水	xin-shui	fuel-water	salary
小说	xiao-shuo	small-talk	novel
主人公	zhu-ren-gong	host-man-sir	heroine
地平线	di-ping-xian	ground-flat-line	horizon
风云人物	fengyun-renwu	weather-figure	celebrity

Table 1: Examples of non-compositional Chinese compounds.

[1] See also (Yun et al., 2002).

[2] A longer discussion about the word formation in general can be found in (Packard, 2000), where a syntactical definition of word according to the X-bar theory is used. In (Lu, 2006) the influence of the segmentation strategy on the definition of what constitutes a word is stressed. An interesting discussion on the headedness of Chinese compounds has been contributed by (Cecaggagno and Scalise, 2006).

that appear in Chinese text in both transliterated and non-transliterated form. In a first pass through the output of the context variety approach, we used a forward-maximum algorithm combined with a set of rules for the detection of transliterated personal names. In a second pass, we used a filter for number-measure-word combinations. Finally, non-transliterated Chinese personal names were removed.

Forward maximum matching (FFM). FFM involves lexicon lookup. From a starting position in the corpus, sequential characters are pushed to a string as long as the new string is a prefix of a lexicon entry (Wong and Chan, 1996). A character that destroys the prefix property marks the recursive starting point of the procedure.[3] In our experiments, a Chinese dictionary with approximately 120,000 entries was used.[4] For efficiency reasons the algorithm is implemented as a cascade.

Transliterated names. In addition to Chinese names, in the wake of globalization, a huge number of foreign names have entered Chinese text. The major part of these names occurs in transliterated form. This means that names, for example, originally represented in Latin script are phonetically transcribed into Chinese characters. In principle, arbitrary Hanzi can be used to transliterate a name. Despite that, conventions seem to exist in that, for example, the name *Mark* is always transliterated with the combination 马克 instead of the phonetically tantamount 马课 or 码克. From a list of 60,000 transliterated English and German names, we extracted the characters that are used for transliteration of syllables. We found a high level of correspondence between the characters used for the transliteration of English names and those for German. Eventually, a transliteration set of 250 Chinese characters was applied to supplement the segmentation algorithm. After the treatment of the numeric compounds, we scan for unexplained strings of Chinese characters of length two that consist only of transliteration Hanzi. If the following character of such a string also is used for transliteration and does not constitute a substring of the base dictionary with its right neighbor, it is added to the transliteration string. By this method, transliterated names, as for example, 安吉利娜茱莉(Angelina Jolie), and also many geographical names, such as, 慕尼黑(Munich) or 弗莱堡(Freiburg), are filtered. On the other hand, some words falling into the introduced class of partially transliterated compounds are lost.

Numeric compounds. A characteristic feature of Chinese is the existence of approximately 200 specialized measure words that combine with different objects and areas (Mo et al., 1996), as for example, 二百万个(two million), 第一届(the first session), or 2007年(year 2007). After the first pass of the segmenter, number-measure-word combinations are filtered from the result set. If a token is a string of numbers we test the following two characters on the measure-word property. The successfully parsed combinations are excluded from the compound set.

Chinese names. The non-transliterated Chinese names are treated after the second segmentation phase: Chinese last names are represented by one or two Hanzi derived from a restricted set of approximately 700 single and 100 double characters, whereas Chinese first names come in principle from an open character set. However, certain characters with a strongly negative sentimental connotation, as for example 死(dead) or 杀(murder), never occur in a Chinese name. Additionally, to this non-name characters, the algorithm distinguishes the treatment of usual and unusual last names. If the algorithm detects a usual last name, the next two tokens of the segmentation result are examined whether they are a potential first name. In case an unusual last name is found, the potential first name is additionally matched against a base dictionary to prevent false positives. All recognized first names are stored to enable a detection independent from the last name later in the text.

3.3. Syntactical step

After the treatment of the context variety output with the augmented segmenter, the major remaining problem causing false positives are complex syntactical structures (phrases) and other than nominal compounds that are mixed with the target entities.

Tagging. With a tagger the part of speech classes of the segmented compound candidates are established. The tagger was implemented as a Hidden Markov Model and trained using the Chinese Treebank of the University of Pennsylvania. As a default tag, we used *NR*, denoting a non-recognized POS. To employ tagging for the recognition of noun-compounds makes sense because the classes of the elements of nominal compounds in Mandarin are not arbitrary. As an example, the possible combinations for a nominal compound consisting of two morphemes fall into the following set: *NN+NN, JJ+NN, NR+NN, NR+NR, NN+NR, DT+NN, CD+NN, NR+JJ, AD+VV*. Probabilities of POS-combinations are used to separate nominal compounds from other, non-relevant strings.

Parsing. An additional improvement of the results could be achieved by uncovering the internal syntactical dependencies of falsely recognized nominal compounds (*pseudo nominal compounds*). In Table 2, we provide examples of compounds and pseudo compounds annotated according to the standard of the Chinese Treebank (Xia, 2000; Santorini, 1990). As the annotation demonstrates, by mere POS tagging both types can not be distinguished. At the constituent level, selective clues emerge: embedded NPs strongly indicate a pseudo compound, i.e., a complex phrasal structure. However, a problem with constituent parsing, so far, is that it has a serious computational overhead and that the accuracy still is problematic.[5] Because of that, we applied an alternative method that tries to derive the selective structural information by a statistical procedure.

[3] The maximum-matching algorithm can be implemented in a forward and a backward variant. Despite the superiority of the backward-method we used forward-maximum, since the diverse filters were much easier to implement

[4] The used lexicon is publicly available at www.mandarintools.com.

[5] Furthermore, the consequences of specific features such as topic-subject constructions as in 这件衬衣颜色不好 that lead to NN combinations at the POS level have to be further investigated (Li, 2005). Less complex syntactical methods, as for example, a dependency parser will be evaluated in a forthcoming paper.

Nominal Compound
(NP (NN 建筑) (NN 公司)) construction company
(NP (NN 经济) (NN 不景气)) recession
(NP (NN 运动)(NN 器材)) sports equipment
Pseudo Nominal Compound
(NP-OBJ (ADJP (JJ 违章)) (NP (NN 建筑))) illegal construction
(NP (DP (DT 全)) (NP (NN 国))) the whole country
(NP (ADJP (JJ 具体)) (NP (NN 措施))) concrete measure

Table 2: Examples of nominal compounds and pseudo nominal compounds.

3.4. Supervised Learning step

To uncover pseudo-compounds, dispensing with syntactical parsing, we used the probability of POS-sequences combined with the pointwise mutual information (MI) between the elements of a compound candidate. MI compares the *joint probability* of seeing two words/morphemes together with the *chance probability* (Church and Hanks, 1989). The underlying hypothesis for our problem is that the elements of a compound are statistically more strongly related than the elements of a phrase. Practically, MI is computed by the the frequency counts of Morphemes in a reference corpus, $f(M_i, M_j), f(M_i), f(M_j)$ normalized by the corpus size N. The pointwise mutual information is: $MI = \frac{f(M_i,M_j)N}{f(M_i)f(M_j)}$. The probability of POS-sequences appearing in compounds as compared to pseudo-compounds was modeled by frequencies of a Treebank training set. The two sources of information were combined by an SVM-classifier implementing pairwise training.[6]

4. Experiments

For the following experiments, we used the Xinhua part of the Chinese Gigaword Corpus (LDC2003T09) with approximately 382 million Chinese characters. We extracted a random sample of 200 *story texts* with 119,509 Hanzi characters.[7] All compound words of this evaluation corpus were tagged, segmented into their morphemes, and augmented with the POS-information of their segments.[8] Recall and precision were measured for types of the collected compounds (i.e., each compound counts only once) as the final goal of the approach is to create a comprehensive dictionary.[9]

DEL	PRE	SUC
了	者	阿
的	子	小
在	们	副
和	术	总

Table 3: Examples of delimiters (DEL), adjacent predecessors (PRE), and adjacent successors (SUC).

4.1. Accessor Variety + Adjacent Character Method

In the first experiment, we applied accessor variety combined with a filtering of adjacent characters to establish compound candidacy. Using a balance parameter, we collected 16 delimiters, 50 adjacent predecessors and 97 adjacent successors.[10] Examples for these grammatical characters are given in Table 3. The results for different values of RV, AV that privilege either precision (first part) or recall (second part) are documented in Table 4. To measure the effects of corpus size, an additional experiment was conducted on the full Xinhua part of the Gigaword corpus with precision and recall again evaluated on our 200 story corpus.

Recall. That the algorithm delivers only a small part of the data is to be expected, since according to Zipf's law most words are infrequent so their context variety falls bellow the threshold. For the bigger corpus (see Table 5) the recall was worse; a part of the effect occurred because of higher RV values that prevented some of the very rare compounds to be collected. Another effect concerned compounds that had the same contexts in the additional material and by that lost relative context variety.

Precision. The precision for the two 2-character compounds is extremely low because many of the recognized candidates, for example 学校(school), are not compounds but simple words. For 3-morpheme compounds the precision soars, but is still insufficient for lexicon construction. Many Chinese names are three-character words that are falsely recognized as compounds. For the 4-character and 5-character compounds the major remaining problem are phrases, which are recognized as compounds. The latter improved significantly for the experiment on the bigger corpus (see Table 5).

4.2. Augmented Segmenter Method

To exclude non-segmentable 2-character units and to separate 3-character Chinese and transliterated foreign names from the compounds, we used the augmented segmentation algorithm, described in Section 3.2.[11] The precision and re-

[6]For the SVM-experiments we applied the WEKA implementation (Witten and Eibe, 2005).

[7]Documents in the Gigaword corpus that are tagged as *story* represent a coherent text on a certain topic.

[8]For research purposes the evaluation corpus is available upon request. The percentages of the 1,624 annotated nominal compounds according to the number of characters are C=2: 13.1%,C=3: 23.1%,C=4: 43.2%,C=5: 20.6%.

[9]In difference to Feng+04, where the tokens of recognized compounds were measured, that is, each compound is counted as often as it appears in the evaluation corpus.

[10]The balance parameter is defined as: $\frac{|RAV-LAV|}{Min(RAV;LAV)}$. Delimiters are expected to be balanced, whereas right and left adjacent characters are unbalanced to the right respectively left context value.

[11]For the implementation of the segmenter, we partially applied code from Eric Peterson, publicly available at

Char	AV	RV	Recall	Precision
2	12	6	4.2%	2.4%
3	10	5	4.0%	38.5%
4	8	4	3.6%	89.3%
5	6	3	5.1%	66.7%
2	10	5	5.1%	2.3%
3	6	3	14.9%	38.4%
4	4	2	17.7%	66.0%
5	4	2	18.3%	51.5%

Table 4: Accessor Variety. Recall and Precision for recognized nominal compounds in the 200 story evaluation corpus.

Char	AV	RV	Recall	Precision
2	18	9	4.5%	2.6%
3	14	7	11.7%	43.2%
4	10	5	12.7%	90.3%
5	7	3	15.4%	68.7%

Table 5: Accessor Variety. Recall and Precision for recognized nominal compounds measured on the 200 story evaluation corpus, with an experimental run on the whole Xinhua Gigaword corpus.

call for Chinese names were 80% and 88%, for transliterated names 67% and 60%. The results for the recognition of compounds on our evaluation corpus are summarized in Table 6. Remaining errors concern other than nominal compounds and phrases.

4.3. POS-Tagging

The POS-tagger, implemented as a HMM and trained with the Chinese Treebank (LDC2005T01), excluded candidates with tag-sequences that do not fall into the set of possible combinations for nominal compounds. By this, phrases and other than nominal compounds were filtered (see Table 7). For example, the four-character string "发表声明" (to give an explanation) leads to the tagging-result "发表(VV) 声明(NN)" which, with high probability, is the signature of a verbal phrase.

4.4. Supervised Learning

With a combination of MI and the probability of a POS-pattern, an SVM-classifier is trained to separate nominal and pseudo-nominal compounds. The final results show further improvement, leading to a very high precision of over 95% (see Table 8).

www.mandarintools.com.

Char	AV	RV	Recall	Precision
2	10	5	4.7%	78.6%
3	6	3	9.9%	62.2%
4	4	2	14.8%	68.9%
5	4	2	8.9%	58.1%

Table 6: Accessor Variety + Segmenter. Recall and Precision for recognized nominal compounds in the 200 story evaluation corpus.

Char	AV	RV	Recall	Precision
2	10	5	2.8%	91.7%
3	6	3	7.9%	87.5%
4	4	2	14.1%	96.9%
5	4	2	7.1%	84.1%

Table 7: Accessor Variety + Segmenter + Tagger. Recall and Precision for recognized nominal compounds in the 200 story evaluation corpus.

Char	AV	RV	Recall	Precision
2	10	5	2.2%	99.0%
3	6	3	6.2%	93.0%
4	4	2	11.1%	98.0%
5	4	2	5.6%	96.0%

Table 8: Accessor Variety + Segmenter + Tagger + SVM Classifier. Recall and Precision for recognized nominal compounds in the 200 story evaluation corpus.

4.5. MI and Compounding

For a last experiment, we augmented the starting accessor variety algorithm for compounds consisting of more than two characters with an MI-based extension. If the pointwise mutual information between the parts of a compound exceeds a certain threshold, it is accepted as a candidate independently of its context variety. The recall for rare compounds increased significantly; but before filtering the precision was far too low for further processing. The initially low precision improved significantly after the application of the filters. A further improvement is to be expected for a tuning of the machine learning classifier with training data that better reflects the distribution of the compounds after the initial filtering.

Char	Recall	Precision	$Recall^F$	$Precision^F$
3	32.3%	12.0%	25.1%	50.7%
4	73.1%	20.0%	69.0%	59.6%
5	56.3%	13.0%	45.1%	44.3%

Table 9: MI as additional criterion to establish compound candidacy. Recall and Precision before and after filtering.

5. Conclusion

In this paper, we introduced and evaluated a pipeline to collect Chinese nominal compounds from arbitrary corpora. A cascade of filters applied to a preliminary set of compound candidates led to a very high precision of over 90%, measured for the types. The result also holds for a small corpus where a solely contextual method introduces too much noise, even for the longer compounds. An introduction of MI into the basic candidacy algorithm led to a much higher

recall with still reasonable precision for subsequent manual processing. In case of the four character compounds, that represent over 40% of the target data, the method has sufficient efficacy to support the rapid construction of compound dictionaries from domain corpora.

Further research aims at the integration of an additional semantic filter that exploits semantic constraints at the character level, as proposed in (Sun et al., 2005) as well as the recognition of descriptive compounds for specific domains.

Acknowledgments

This work is supported by iCORE, AICML, and Google.

6. References

Antonella Cecaggagno and Sergio Scalise. 2006. Classification, structure and headedness of chinese compounds. *Lingue Linguaggio*, V.2:233–260.

Kenneth W. Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, B.C. Association for Computational Linguistics.

Haodi Feng, Xiaotie Deng, Kang Chen, and Weimin Zheng. 2004. Accessor variety criteria for Chinese word extraction. *Computational Linguistics*, 30(1):75–93.

Franz Guenthner and Xavier Blanco. 2004. Multi-lexemic expressions: an overview. *Lingvisticae Investigationes Suplementa*, 24:239–252.

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

Jack Halpern. 2006. The role of lexical resources in cjk natural language processing. In *Proceedings of the workshop of Multilingual Language Resources and Interoperability (MLRI06)*.

Saul Hansel. June, 3rd, 2007. Google keeps tweaking its search engine. *The New York Times*.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Comput. Linguist.*, 29(3):333–347.

Charles N. Li and Sandra A. Thompson. 1989. *Mandarin Chinese - A Functional Reference Grammar*. University of California Press, London, United Kingdom.

Wendan Li. 2005. *Topic Chains in Chinese*. Lincom Studies in Asian Linguistics, Munich, Germany.

Xiaofei Lu. 2006. *Hybrid Models for Chinese Unknown Word Resolution*. Ph.D. thesis, Ohio State University.

Igor Melcuk. 1995. The future of the lexicon in linguistic description and the explanatory combinatorial dictionary. In I.-H. Lee, editor, *Linguistics in the Morning Calm 3 (Selected Papers from SICOL-1992)*, pages 181–270. Seoul.

Ruo-Ping J. Mo, Yao-Jung Yang, Keh-Jiann Chen, and Chu-Ren Huang. 1996. Determinative-measure compounds in mandarin chinsese: Formation rules and parser implementation. In C.R.Huang, K.J.Chen, and B.K. Tsou, editors, *Readings in Chinese Natural Language Porcessing*, pages 123–146.

Jerome Packard. 2000. *The Morphology of Chinese: A Linguistics and Cognitive Approach*. Cambridge University Press, Cambridge.

Christoph Ringlstetter, Klaus U. Schulz, and Stoyan Mihov. 2007. Adaptive Text Correction with Web-Crawled Domain-Dependent Dictionaries. *ACM Transactions on Speech and Language Processing*, 4(4):(9) 1–36.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *CICLing*, pages 1–15.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the penn treebank project. Technical Report Technical report MS-CIS-90–47, Department of Computer and Information Science, University of Pennsylvania.

Stanley Starosta, Koenraad Kuiper, Siew-Ai Ng, and Zhi-Quian Wu. 1998. On defining the Chinese compound word: Headedness in Chinese compounding and Chinese vr compounds. In Jerome L. Packard, editor, *New Approaches to Chinese Word Formation: Morphology, phonology and the lexicon in modern and ancient Chinese*, Trends in Linguistics. Studies and Monographs, pages 347–370. Mouton-de Gruyter, Berlin.

Maosong Sun, Shengfen Luo, and Benjamin K T'sou. 2005. Word extraction based on semantic constraints in chinese word-formation. In *Proceedings of CICLING 2005, LNCS 3406*, pages 202–213.

Evelyne Viegas, Wanying Jin, Ron Dolan, and Stephen Beale. 1998. Representation and processing of Chinese nominals and compounds. In *Proceedings of the Workshop on Content Visualization and Intermedia Representations (CVIR'98) of the 17th International Conference on Computational Linguistics (COLING '98) and the 36th Annual Meeting of the Association for Computational Linguistics (ACL'98)*, August.

Seretan Violeta, Luka Nerima, and Eric Wehrli. 2004. A Tool for Multi-Word Collocation Extraction and Visualization in Multilingual Corpora. In *In Proceedings of the Eleventh EURALEX International Congress (EURALEX 2004)*, pages 755–766.

Ian H. Witten and Frank Eibe. 2005. Data mining: practical machine learning tools and techniques. 2nd edition. Morgan Kaufmann, San Francisco. http://www.cs.waikato.ac.nz/ ml/weka.

P. Wong and C. Chan. 1996. Chinese word segmentation based on maximum matching and word binding force.

Fei Xia. 2000. Segmentation guideline for the chinese treebank project. Technical report, Department of Computer and Information Science, University of Pennsylvania.

Liu Yun, Yu Shiwen, and Zhu Xuefeng. 2002. Construction of the contemporary chinese compound words database and its application. In Chinese.

Weidong Zhan. 2000. Construction of Words: A Course of Contemporary Chinese. *Tsinghua University Press*.