Distributional Evidence and Beyond: the Success and Limitations of Machine Learning in Chinese Word Segmentation

Jianqiang Ma
Department of Linguistics
University of Tübingen
Wilhelmstr. 19, Tübingen, 72074,
Germany
jma@sfs.uni-tuebingen.de

Dale Gerdemann
Department of Linguistics
University of Tübingen
Wilhelmstr. 19, Tübingen, 72074,
Germany
dg@sfs.uni-tuebingen.de

Abstract

In this paper, we argue that the key to the success of the current state-of-art statistical learning algorithms for Chinese word segmentation (CWS) mostly lies in their optimal weighting of non-overlapping *distributional evidence* in the corpora. The utilization of distributional evidence is more essential than the learning algorithm. We further analyze the characteristics of distributional evidence for CWS, under the framework of Zipf's law and summarize the limitation of statistical learning in CWS as the *feature absence problem*, which may be apparent yet usually neglected. Making a connection between theoretical/empirical linguistics and CWS, we suggest that the study and development of a *generative word formation system* may be beneficial for both the science and engineering of CWS. We wrap up the discussion after reviewing some recent works that are already on this line.

1 Introduction

Tokens in general, words are considered as building blocks of linguistic structures of human languages and basic inputs for natural language processing (Webster and Kit 1992). In many Asian languages, including Chinese, sentences are written as character sequences without explicit word delimiters, thus tokenization or *word segmentation* remains a key research topic in language processing for these languages.

The most popular model among modern word segmenters is probably character position tagging (Xue, 2003), which views word segmentation as labeling the positional roles that character plays within words, using labels such as **B**eginning, **M**iddle, **E**nding and **S**ingleton. Under such formulation, Chinese word segmentation (CWS) becomes a special case of sequence labeling problem, which can be effectively solved by machine learning techniques such as conditional random fields (Lafferty et al., 2001), which achieves state-of-art results for CWS.

In recent years, the performance of machine learning based segmenters has been further pushed forward by model combination (Wang et al., 2010; Sun, 2010), utilizing unsupervised segmentation on unlabelled data (Zhao and Kit, 2008; Sun and Jia, 2011), jointly learning segmentation and POS tagging (Jiang et al., 2008; Zhang and Clark, 2008; Sun, 2011), etc. On the other hand, it appears that out-of-vocabulary words (OOV) remain a major challenge even for these sophisticated systems. Given this background, our paper attempts to analyze both successes and limitations of machine learning approaches to CWS, in the hope of bringing new understandings and inspiring novel methods.

First of all, what types of *evidence* (information/feature) are most important for any segmenter? The most intuitive choice is *lexical forms*, which have been extensively used by early systems in the form of dictionary or token functions (e.g. frequency). However, as lexical forms are incapable of describing morphological behaviors of characters, it fails to contribute to recognition of OOV, which exist as a result of dynamic and productive word formation in Chinese. It turns out that character information alone provide adequate information for describing both IV (in-vocabulary words) and OOV, suggested by the success of various character position tagging systems. Specifically, such systems mostly rely on *character distributional evidence*, i.e. characters and character co-occurrences in different positions of words or word sequences.

Another important question is what role machine learning algorithms play. It might seem that the machine learning algorithm is a black box where magic happens, i.e. machine learning should get all the credit for the improvement over the well established baseline of maximum matching (Liang, 1986). But this needs more careful examination. We show in section 3 that the role of machine learning in CWS systems can be better described as *feature weight optimization*.

One implication of above mentioned issues is that despite different strategies for feature weight optimization, the performance of virtually *all* the current machine learning based segmenter are bounded by what can be expressed by character distributional evidence. Like many other linguistics phenomena, the character ngram distribution is characterized by *Zipf's law* (Zipf, 1949), which states that relatively few items are very frequent while most items are rare. Given Zipf's law, the distributional features that we have acquired from the training corpus are likely to cover only a subset of distributional features of the testing corpus, as some of rare features may only appear in either corpus but *not* both. This is consistent with our empirical study of distributional evidence and is exactly the problem for recognizing OOV. So the real challenge in CWS is that the distributional evidence for some characters in OOV is at least partly unavailable, where algorithmic predictions yield only low accurate guessing.

Similar to the limitation of machine learning in CWS, Yang (2011) suggests that usage/item-based theory in language acquisition (Tomasello, 2000; Hay and Baayen, 2005) has drawbacks on modeling the empirical data, also because of the Zipf's law. The generative linguistic system, on the other hand, is consistent with the language acquisition data. Interestingly, recent development of Chinese morphology, such as Packard (2000) and Xue (2001) also argues that it is attractive to describe word formation in Chinese using generative rules with part-of-speech like tags. These theoretical advancements shed light on new paths to solving the OOV problem in word segmentation with generative word formation models. Our discussion finishes by summarizing some pilot work that are already in this direction, including work of the current authors.

2 Distributional Evidence for CWS

Early work in CWS extensively use lexical forms as the main information source. In maximum matching, sentence substrings that match lexical entries in the dictionary are selected as word candidates and the disambiguation of conflicting segmentations are achieved in a greedy search way. In finite state methods for CWS such as Sproat et al. (1996), lexicon is represented as weighted finite state machines and the segmentation disambiguation is based on scores of individual lexical item given by the finite state machine, which is mostly trained from word frequency statistics. But the rise of character position tagging approach to CWS shows that the lexical information is *neither necessary nor adequate* for the building accurate CWS systems. On one hand, various systems mainly using character distribution information (Xue, 2003; Peng et al., 2004) have similar results on IVs compared with word-based systems. On the other hand, character position tagging systems have very strong power on OOV recognition, which word-based systems basically fail to do.

Note that even for recent discriminative learning powered word-based segmenters (Zhang and Clark, 2008) that have state-of-art performances, character level features have been widely integrated. Actually, while it is hard to imagine how OOV can be properly modeled if all the character information is removed, discarding all lexical information may just end up with a system somewhat similar to a character tagging system.

2.1 Character Features that matter

In fact, lexical forms can be viewed as a special case of character distributional information, as the lexicon is a set of character sequences (co-occurrences). Some of the most useful character features proposed in Xue (2003) are following:

- Character unigrams: C_s (i-2<s<i+2)
- Character bigrams: $C_s\,C_{s+1}$ (i-2<s<i+2)
- Tag unigrams: T_s (s=i-1, i-2)

, where C represents a character, T represents a tag, s denotes the position index of the character string and i denotes the position of the current character of interest.

It can be seen that besides the interactions with character position tags, features are basically *character co-occurrences*. This feature set has been widely adopted in many latter systems, complemented by extra features such as punctuation, date, digit and letter, tone, etc. (Zhao et al., 2006). But it is fair to say the improvement brought by extra features is useful yet marginal.

2.2 The minority rules

It is not adequate to know that character distributional evidence is the dominant information for segmentation. As features do not necessarily contribute equally to the task, it is more interesting to examine *how* different features influence the segmentation. Feature pruning provides a good perspective to understanding the contributions of individual features. If some features are pruned without significantly hurting the performance, these features may be less crucial or at least redundant with regard to the remaining features. Zhao & Kit (2009) have proposed a simple and efficient model pruning method for conditional random fields. A closer look at their experiments results on CRF based CWS helps us better understand the roles that different features play. The general message is that standard features as mentioned in previous section are *highly redundant*. According to their report, the model that uses only 2% of total number of features that have survived the pruning process can still reach above 97% of the accuracy of that which can be accomplished with the full feature set. Moreover, no performance loss occurs at all until the pruning rate is larger than 65%. In other words, a few features contribute a great deal to the performance of the current state-of-the art system.

Researchers have also found similar patterns on other sequence labeling tasks such as named entity recognition and chunking as well (Goldberg & Elhadad, 2009). It has shown that accurate models for these tasks can be learned from a heavily pruned feature space, which contains less than 1% of the features in the training set. In their experiments it turns out that rare features are used for ruling out uncertain cases by the machine learning algorithm rather than learning useful generalizations. We speculate that this conclusion might also be true for CWS task and we will further discuss the characteristics of the distribution of character ngram features in section 4.

3 The Role of Machine Learning

The Chinese language processing community has witnessed a dramatic performance boom of CWS systems since the introduction of machine learning algorithms under the character position tagging framework. It appears that machine learning is the black box where magic happens, as there is a huge gap between the state-of-the-art machine learning systems and the traditional dictionary-based greedy search baseline maximum matching. However, since most machine learning based systems dominantly rely on character distributional evidence, one may wonder whether the character distributional evidence within the framework of character position tagging should be given more credit that they have deserved. Our preliminary study has also shown that it is possible to achieve more than half of the error reduction on OOV recognition that the-state-of-art methods can achieve, by a simple combination of distributional evidence.

The re-examination of the role of machine learning in CWS is useful for a thorough understanding on *how* machine learning contributes to this task. To simplify the discussion, we restrict ourselves to log linear family of learning algorithms, i.e. maximum entropy, CRF, etc. These algorithms combine the features in a linear way and the learning process is reduced to the estimation of feature weights. But the argument should also hold for other algorithms, such as artificial neural networks, the only difference of which in this context is that there are hidden nodes that represents non-linear combinations of features. In either case, what machine learning *can* do is to optimize the weights for features using different strategies. Thus the role of machine learning can be summarized as feature weight optimization. This understanding is important as one should distinguish the challenge in optimization for a given the feature space and the inherent problems of feature space itself. As we will show later sections, this links closely to the limitations of machine learning approaches to CWS and calls for new perspective of looking at CWS.

4 The Zipfian Distribution of Distributional Evidence

4.1 The feature absence problem of OOVs

OOVs are considered to be the major error source in the state-of-the-art machine learning based CWS systems. While those systems can achieve accuracy (F-score) over 95% on treebank corpora, their recall on OOVs are typically only around 70% (Emerson, 2005; Levow, 2006; Zhao and Liu, 2010). In order to illustrate the main problems of machine learning approaches to CWS, we have conducted an empirical study on those OOVs that the modern CWS systems fail to recognize. We are particularly interested in whether those errors are caused by feature weight optimization problems, or the inherent problems of the feature space itself.

The study is based on Penn Chinese Treebank version 5 (Xue et al., 2005), which is manually word-segmented. We trained a CRF based segmenter on 75% of the corpus and use the model to segment the remaining 25%. Those words only occur in the training section but not the testing section are considered as OOVs. The OOV rate is about 9% in this set-up. We use a simplified version of feature template proposed in Xue (2003) for training, namely only current characters (C_0), current and previous characters ($C_{-1}C_0$, denoted as B_1) as well as current and next characters (C_0C_1, denoted as B_2), i.e. unigrams and left/right bigrams. This choice is for the purpose of concentrating on the dominant factors and simplifying the discussion, given the fact that those features contribute more than 98% of the overall accuracy and 95% of OOV recall on this corpus.

One observation about those error-causing OOVs has drawn our attention. Among all character instances,

- 1.6% have C_0 feature unseen and thus B_1 and B_2 feature unseen in the training corpus (Type I);
- 29.2% have only C_0 feature seen, but both B_1 and B_2 features are unseen (Type II);
- 36.2% have and only have one of the bigram features unseen, i.e. either B_1 and B_2 is unseen (Type III).

In other words, 67% of character instances have at least one of the features B_1 and B_2 unseen from the training corpus, while only 23% of character instances have both B_1 and B_2 seen in the training corpus.

We may call this phenomenon as the *feature absence problem*. Type I is apparently fatal for any meaningful prediction, as there is not *any* feature at all for the model to utilize. Type II is also disastrous for a sensible prediction, as the unigram feature C0 alone could hardly determine the label or the role of the character correctly. In Chinese, the majority of character may occur in any position of a word, i.e. its label can be either **S**tart, **M**iddle, **E**nd or **S**ingleton, except for a few characters which have dominant roles such as prefix (e.g. 非, 'not/non', 反 'anti') or suffix (e.g. 者 'one who does or is ...', 化 a verbalizing suffix). Note that even for these characters, there are ambiguities as for the role in a word, e.g. 非 can be the end of a word as in 是非 'right and wrong/quarrel'.

Character instances in Type III have a better chance of being correctly labeled by the model but relying only on the bigram context on one side is likely to be of high bias in the first place, and it might be the case that the bigram context on "the other side" is more informative than the one that are seen in the training corpus. Moreover, the association of a certain character co-occurrence with a certain label in the training corpus might also be merely by chance, especially for those co-occurrences that are less frequent in the training corpus. Finally, the statistics here is with regard to characters, and we need be aware that the recognition of an OOV fails even if only one the character is incorrectly labeled, which means this 67% feature absence case may explain a much higher percentage of OOV tokens that are not recognized.

It is clear that the issue above is an inherent problem of the feature space and is out of reach for the clever optimizations offered by machine learning algorithms. To illustrate this, we fit the discussion in an abstract view of classification algorithms in machine learning. The model can be viewed as hyper planes that separate the feature space, in which the training instances are dots. The separation should be made in such as a way that instances of the same class are in the same subspace, if noise are not taken into account. The prediction or testing process is fairly straightforward once these hyper planes are determined in the training process. For an new/unseen instance, its features corresponds to coordinates of dimensions in the space, once the coordinates are determined, the instance fits an area, preferably a dot, in the space separated by the model. The subspace that the instance falls in defines its label. However, the situation in the feature absence problem is that very few, or in extreme cases, no coordinates are given for the new instance in testing data, thus the area in the space determined by these coordinates are so vast that they may cross the boundaries of the hyper planes. In this case, one would not be able to tell which subspace or class that instance belongs to. Of course, sequence labeling is more complicated than classification, but the above argument also holds.

4.2 Zipf’s law and its implications

The problem seems to be that our training corpus is too small to contain all the bigram co-occurrences that occur in the testing corpus. So can we simply enlarge our training corpus to solve this problem? Unfortunately, there are two factors that make this proposal less appealing as at the first glance. Firstly, the training corpus is obtained via human annotations, which are expensive. Secondly, empirical study shows that the scale of corpus that we need to capture enough features grows at an exponential rate with regard to the number of distinct features (Zhao et al., 2010). The second factor is determined by the Zipf’s law (Zipf, 1949), which widely applicable to linguistics data and empirical distributions in many other areas.

Zipf’s law states that the frequency of an item (character, word, bigram, etc.) is approximately equal to the inverse of its rank in frequency, which can be expressed by the following formula:

$$f = C/r \tag{1}$$

, where C is some constant, f is the frequency of the item and r is its rank of frequency in the set of the item. A perfect Zipfian distribution would be a straight line of slope -1, with the axes being log of word frequency and the log of word rank. The empirical usually have minor deviation from the perfect scenario (Figure 1). There are many vocabulary studies that report Zipf’s law in various language and genres (Baroni, 2008). The distribution of Chinese characters ngrams and word ngrams are of no exception.

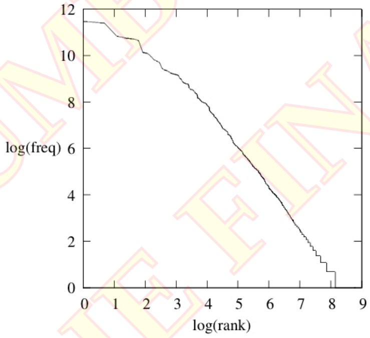

Figure 1: A Zifp’s law curve of word frequency

One immediate implication is that only a small percentage of items occur very frequently while the majority of items occur very rarely (in extreme but common cases, the frequency equals to 1) in real texts. On the word level, Zipf’s law suggests that given a relatively small sample of the sentences in a language such as a treebank corpus, most words are so rare that they are likely to occur in either training section or testing section but *not* both, which explains why OOVs occur in the first place.

On the character level, Zipf's law means that most character co-occurrences are so rare that they occur either in training or testing corpus but not both and only a few character co-occurrences are so frequent that they are likely to occur in both corpus. This explains why very heavy feature pruning works, as a small subset of all character co-occurrence types take account for the most co-occurrence tokens. In short, the feature absence problem is a rule rather than an exception.

Another interpretation of Zipf's law is that it predicts how large a corpus is needed to cover a certain number of distinct word/character ngrams. The general idea is that since a few items occur very frequently, a new non-frequent item will only appear after seeing many occurrences of these frequent items. Mathematically, the sum of all relative frequencies in a Zipf distribution is equal to the harmonic series and therefore:

$$\sum_{n=1}^{\infty} \frac{1}{n} = \infty. \qquad (2)$$

This formula states that series gets arbitrarily large as *n* becomes larger, which suggests that exponential more tokens have to occur before more distinct types are encountered. This has been confirmed by empirical study (Zhao et al., 2010) as well. The bad news is that even though the scale of commonly seen Chinese characters is only at thousands. The word formation process that combines characters is very dynamic and productive. Even if we only consider words that are made of two characters, the upper bound of number of extinct types is 10^6 ($10^3 \times 10^3$). Although the actual number of distinct two-character words is far smaller than the upper bound, the scale of annotated corpus needed to solve the data sparseness problem is still tremendous. Given the inevitable presence of the feature absence problem, which is governed by Zipf's law, the efforts on solving the OOV recognition problem by applying stronger machine learning algorithms or smarter system combination are beneficial yet seem to aim only at the tip of the iceberg.

5 Relevance to Language Acquisition

Before we move to the discussion of possible solutions of the OOV recognition problem, let us first examine an interesting connection between the limitation of feature-based machine learning approaches to CWS and the drawbacks of the item-based approach to language acquisition.

Since Chomsky (1965), linguists have been aware of the distinction between competence and performance, which suggests that it is limited to draw conclusions only from observed linguistic data. For example, some words have never been said but are nevertheless grammatically correct. This distinction has also been widely accepted in the subfield of language acquisition, even by researchers that do not follow the generative grammar. However, this idea has been recently challenged by the *item or usage based* *theory* of language acquisition (Tomasello, 2000; Hay & Baayen 2005, etc). The item-based approach states that language acquisition can be achieved by memorizing and operating specific schemas of linguistic forms and constructions, in contrast with the traditional thought of learning grammar rules that consist a productive/generative linguistic system.

Note that the claim of the item-based approach to language acquisition is similar to the feature-based machine learning approaches to CWS at an abstract level. Both approaches build models using specific surface linguistic forms and their co-occurrences and the model retrieves such stored "pairings of form and functions" to do the production or recognition, although the models in the former do not necessarily of statistical nature as those in the latter do.

Interestingly, the generative school fights back (Young 2011) and argues that there are some inherent limitations in the item-based approach, as Zipfian distribution determines that most "pairings of form and functions" will never be heard and even for those do occur may be so infrequent that the storage of usage of such pairings is not reliable. Further empirical study has shown that the item- based approach is not supported by statistical evidence in language acquisition data. On the contrary, generative grammars are consistent with empirical data, based on a model that considers the interaction of Zipfian distribution and the combinations of linguistic items.

While CWS is a different domain than language acquisition, the arguments here may still provide a hint on understanding the OOV problem. It is likely that the Zipfian nature of character/word ngram distributions ensures that the overlap of these surface form co-occurrence based features in training and testing corpus of CWS systems are quite low by type unless the corpus size is very large, which unfortunately requires an exponential growth of the size of the annotated corpus. And the consistency of empirical data with generative grammars that have been observed in language acquisition case studies may also hold in the word formation process of Chinese, which implies an alternative formalism for solving CWS problem in general and OOV problem in particular.

6 Generative Word Formation Model

The idea that word formation in Chinese is an generative system is reasonable in both language acquisition and theoretical linguistics. This Morphology of Chinese, which is represented by early works such as (Zhao, 1968; Lü, 1979) and more recent work in the framework of generative linguistics such as (Huang, 1984; Dai, 1992; Duanmu, 1997; Packard, 2000; Xue, 2001).

Dai (1992) introduced the idea that different notations of wordhood co-exist, including morphological word, syntactic word and phonological word. The interactions between them explain various word formation phenomena. But his model is basically a static lexicon, which does not provide a concrete proposal on how morphological words are derived.

Packard (2000) is probably the most influential modern work, which treats the morphology as an extension of syntax below the word (X_0) level, following the thinking of Selkirk (1982). Packard (2000) is based on the "form class description", which assigns words and their components (characters) part-of-speech like tags called form class. He has also suggested so called "Headness Principle", which states that nouns have nominal components (characters) on the right and verbs have verbal components (characters) on the left. Like Dai (1992), Packard (2000) also fits into a lexicalism framework, and considers both morphemes and complex words with their "precompiled" morphological structures in the lexicon, except for complex words containing grammatical affixes.

In contrast, Xue (2001) have proposed a system that derives virtually *all* the complex words *using syntax rules* or in the morphology module after syntactic analysis, following the theory of distributed morphology (Halle & Marantz 1993, 1994). The boundary of syntax and morphology further blurred and the operation scope of syntax rules expand to most parts of the morphology.

Despite the disagreements, both Packard (2000) and Xue (2001) agree that part-of-speech like tags for characters and words and syntactic or morphological rules that describe the derivation of these tags make essential parts of a *generative word formation system* for Chinese. Computational linguists have started rethinking the limitations of feature based machine learning approach for CWS and has called for morphology-based analysis of OOVs (Dong et al., 2010). Furthermore, there are already pilot works in this direction, such as Zhao (2009), Li (2011) and Ma et al. (2012). Both methods happen to be formulated as learning a joint model for segmentation and parsing, which has certain practical advantages, but is not necessary for learning a word formation model.

Zhao (2009) has proposed a character-based dependency parsing model, in which the word formation is formulated as the in-word character dependencies, without any part-of-speech tags or dependency labels. The dependency model has comparable performance on the CWS task as the state-of-the art sequence labeling based segmenters. While it is an interesting investigation, pure character-wise dependencies seem to be inadequate to model the word formation process in a general and productive manner.

Li (2011) has proposed a unified parsing model that can parse both word structures and phrase structures. Part-of-speech tags and constituent labels are utilized in this model. The model extends probabilistic context free grammar based constituent parsing to handle the inner structure of words, which has a flavor of generative word formation model, i.e. syntactic rules are used to analyze the word formation process. The performance of this model on CWS task is slightly better than the state-of-the-art but no significant improvement on OOV recognition has been reported. Note that this work makes a distinction between flat words and non-flat words and the grammar model only deals with the generation of the non-flat words. Here the non-flat words are defined as those words that contain productive suffix and/or prefix, which is only a small subset of words that can be possibly analyzed by syntactic or morphological rules. In this sense, Li (2011) can be viewed as an implementation of Packard (2000). The model's low coverage of the word formation phenomena may explain why this model has not brought advancement on OOV recognition. The morphological model might be more powerful on OOV recognition, if syntax-like rules were used to analyze most of, rather than a small portion of, complex words, i.e. by implementing Xue (2001). Nevertheless, the results presented in Li (2011) are encouraging, as it has shown the effectiveness of analyzing word formation using generative rules. Note that Li (2011) follows a standard paradigm in modern syntactic parsing: the probabilistic syntax model that is used for parsing is learned from an annotated treebank. So far, we have also limited our discussion to this default.

Ma et al. (2012) have proposed a semi-automatic approach to Chinese word structure annotation. They have argued that Li (2011) only annotated affixations, which only covered 35% of word types in the corpus and was insufficient to deal with the OOV problem. In contrast, their annotation has covered more morphological phenomena, including compounding, which is a more popular word formation process in Chinese. Unfortunate, the usefulness of such annotation for the OOV problem has not been validated by experiments yet.

One may wonder whether it is possible to have such a strong machine learning algorithm that can overcome the limitations of current learning algorithms used in CWS and effectively induce the word structure without the explicit notion of word formation model and the utilization of manual treebank annotation. This turns out to be quite a difficult task, and the current computational learning research under the framework of Probably Approximately Correct (PAC, Valiant, 1984) suggests that it is virtually impossible to learn languages such as finite state and context free language, given only distribution of surface forms (Yang, 2011). But learnability results are in a general sense and can be modified, e.g. adding certain assumptions, to suit various learning scenario, which is an interesting topic itself.

7 Conclusion

In this paper, we have reviewed some state-of-art methods for Chinese word segmentation, with a focus on the role of distributional evidence and feature-based machine learning algorithms. By showing the Zipfian nature of the distributional evidence, we have further investigated the limitations of feature-based statistical machine learning models for CWS, which can be summarized as the feature absence problem. Drawing the connection with language acquisition literature, we have speculated that a generative linguistic system may help overcome the limitations of current methods. This speculation is supported by some formal linguistic analysis of Chinese morphology. Finally, we have shown that recent results in relevant computational modeling suggests that it is indeed a promising direction to investigate generative word formation models in order to come up with better CWS system.

References

Marco Baroni. 2008. Distributions in text. In Lüdelign, A. & Kytö, M. (Eds.) *Corpus linguistics: An international hanbook*. Mouton de Gruyter, Berlin, Germany.

Noam Chomsky. 1965. Aspects of the theory of syntax. MIT Press, Cambridge, USA.

Xiang-Ling Dai. 1992. *Chinese Morphology and its Interface with the Syntax*. PhD Dissertation, Ohio State University.

Zhendong Dong, Qiang Dong and Changling Hao. 2010. Word segmentation needs change - from a linguist's view. In Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing. Beijing, China.

San Duanmu. 1997. "Wordhood in Chinese", in Jerome J. Packard ed. *New Approaches to Chinese Word Formation*. Mouton de Gruyter, New York, USA.

Thomas Emerson. 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of Forth SIGHAN Workshop on Chinese Language Processing*. Jeju Island, Korea.

Yoav Goldberg and Michael Elhadad. 2009. On the role of lexical features in sequence labeling. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore.

Morris Halle and Alec Marantz. 1993. Distributed Morphology and the pieces of inflection, in Hale, Kenneth and Samuel Jay Keyser eds. *The View from Building 20*. The MIT Press, Cambridge, USA.

Morris Halle and Alec Marantz. 1994. Some key features of Distributed Morphology. *MIT Working Papers in Linguistics* 21, 275-288.

Jennifer Hay and Harald Baayen. 2005. Shifting paradigms: gradient structure in morphology. Trends in Cognitive Sciences, 9, 342-348

James C. T. Huang. 1984. Phrase structure, lexical integrity, and Chinese compounds. *Journal of the Chinese Language Teachers Association* 19.2:53-78.

Wenbin Jiang, Liang Huang, Qun Liu, Yajuan Lu. 2008. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. In *Proceedings of ACL 2008: HLT*. Columbus, USA.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*. Williamstown, MA, USA

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Sydney, Australia

Zhongguo Li. 2011. Parsing the internal structure of words: a new paradigm for Chinese word segmentation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA.

Nanyuan Liang. 1986. On computer automatic word segmentation of written Chinese. *Journal of Chinese Information Processing*, 1(1).

Shuxiang Lü. 1979. *Hanyu Yufa Fenxi Wenti "Problems in the analysis of Chinese grammar"*. Shangwu Yinshuguan, Beijing, China.

Jianqiang Ma, Chunyu Kit and Dale Gerdemann. 2012. Semi-automatic annotation of Chinese word structure. In the Proceedings of *2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing*, Tianjin, China.

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

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of COLING*. Geneva, Switzerland.

Elisabeth O. Selkirk. 1982. The Syntax of Words. Cambridge, Massachusetts: The MIT Press, Cambridge, USA.

Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 22(3):377-404.

Weiwei Sun. 2010. Word-based and character-based word segmentation models: Comparison and combination. In *Proceedings of the 23rd International Conference on Computational Linguistics*: *Posters Session*. Beijing, China.

Weiwei Sun. 2011. A stacked sub-word model for joint Chinese word segmentation and part-of-speech tagging. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, USA.

Weiwei Sun and Jia Xu. 2011. Enhancing Chinese word segmentation using unlabeled data. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK.

Michael Tomasello. 2000. First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11, 61-82.

Leslie G. Valiant. 1984. A theory of the learnable. *Communications of the ACM*, 27, 1134-1142.

Kun Wang, Chengqing Zong and Keh-Yih Su. 2010. A character-based joint model for Chinese word segmentation. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China.

Jonathan Webster and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In *Proceedings of the 14th conference on*. Nates, France.

Nianwen Xue. 2001. Defining and automatically identifying words in Chinese. Phd Thesis, University of Delaware.

Nianwen Xue. 2003. Chinese Word Segmentation as Characater Tagging. *Computational Linguistics and Chinese Language Processing*, 8(1): 29-48

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2)207-238.

Charles Yang. 2011. A computational models of syntactic acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science.*

Yue Zhang and Stephen Clark. 2008. Joint word segmentation and POS tagging using a single perceptron. In *Proceedings of ACL-08: HLT*. Columbus, USA.

Hai Zhao. 2009. Character-level dependencies in Chinese: usefulness and learning. In *Proceedings of the 12th Conference of the European Chapter of the ACL.* Athens, Greece.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006. An Improved Chinese Word Segmentation System with Conditional Random Field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Sydney, Australia.

Hai Zhao and Chunyu Kit. 2008. Unsupervised segmentation helps supervised learning of word segmentation and named entity recognition. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing* (SIGHAN-6). Hyderabad, India.

Hai Zhao and Chunyu Kit. 2009. A simple and efficient model pruning method for conditional random fields. In *Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*. Springer-Verlag Berlin, Heidelberg, Germany.

Hai Zhao, Yan Song and Chunyu Kit. 2010. How Large a Corpus do We Need:Statistical Method vs. Rule-based Method. In *Proceedings of LREC-2010*. Malta.

Hongmei Zhao and Qun Liu. 2010. The CIPS-SIGHAN CLP 2010 Chinese Word Segmentation Bakeoff. In *Proceedings of the First CPS-SIGHAN Joint Conference on Chinese Language Processing*. Beijing, China.

Yuen-Ren Zhao. 1968. *Grammar of Spoken Chinese*. University of California Press, Berkeley and Los Angeles, USA.

George Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wisley. Oxford, UK