Information Sciences 181 (2011) 163–183

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Integrating unsupervised and supervised word segmentation: The role of goodness measures ☆

Hai Zhao [a,b,1], Chunyu Kit [a,*]

[a] Department of Chinese, Translation and Linguistics, City University of Hong Kong, 83 Tat Chee Ave., Kowloon, Hong Kong SAR, PR China
[b] Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, PR China

ARTICLE INFO

Article history:
Received 21 October 2008
Received in revised form 28 April 2010
Accepted 3 September 2010

Keywords:
Chinese word segmentation
Unsupervised segmentation
Unknown word detection
Conditional random fields
Character tagging
Description length gain
Accessor variety
Boundary entropy

ABSTRACT

This study explores the feasibility of integrating unsupervised and supervised segmentation of Chinese texts for enhancing performance beyond the present state-of-the art, focusing on the critical role of the former in enhancing the latter. Following only a pre-defined goodness measure, unsupervised segmentation has the advantage of discovering many new words in raw texts, but it has the disadvantage of inevitably corrupting many known. By contrast, supervised segmentation conventionally trained only on a pre-segmented corpus is particularly good at identifying known words but possesses little intrinsic mechanism to deal with unseen ones until it is formulated as character tagging. To combine their strengths, we empirically evaluate a set of goodness measures, among which description length gain excels in word discovery, but simple strategies like word candidate pruning and assemble segmentation can further improve it. Interestingly, however, accessor variety and boundary entropy, two other goodness measures, are found more effective in enhancing the supervised learning of character tagging with the conditional random fields model. All goodness scores are discretized into feature values to enrich this model. The success of this approach has been verified by our experiments on the benchmark data sets of the last two Bakeoffs: on average, it achieves an error reduction of 6.39% over the best performance of closed test in Bakeoff-3 and ranks first in all five closed test tracks in Bakeoff-4, outperforming other participants significantly and consistently by an error reduction of 8.96%.

1. Introduction

Word segmentation is a primary task for computer processing of many East Asian languages including Chinese, because in the written forms of these languages word boundaries are not explicitly marked with any overt delimiters such as white space. Written Chinese appears in the form of sequence of characters rather than words. Thus, the task of word segmentation is to convert a text as a sequence of consecutive characters into a sequence of correctly delimited words. It is a special case of tokenization for language processing that involves more complicated problems than one would think at the first glance

☆ The research described in this paper was partially supported by the City University of Hong Kong through the Strategic Research Grants (SRG) 7002037, 7002388 and 7008003 and also by the National Natural Science Foundation of China (NSFC) through the Grand 60903119. The progressive results obtained at various stages of this research were presented disjointedly in a number of conferences and workshops [22,54–57]. Heartfelt thanks are given to the two anonymous reviewers for their insightful comments and advice that have helped improve this article significantly, and also to Olivia Kwong and Lisa Raphals for their helps.

* Corresponding author. Tel.: +852 27889310; fax: +852 27887320.
E-mail addresses: zhaohai@cs.sjtu.edu.cn (H. Zhao), ctckit@cityu.edu.hk (C. Kit).

[1] Supported by a postdoc research fellowship in the Department of Chinese, Translation and Linguistics, City University of Hong Kong.

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.09.008

[47,11,30,29]. Various approaches have been explored by scholars to tackle the two main causes of segmentation errors, namely, segmentation ambiguities and unknown (or out-of-vocabulary, OOV) words. Numerous efforts have also been devoted to examine and maximize the effectiveness of various kinds of language resource and linguistics information. For example, the critical role that morphemes and part-of-speeches can play is examined via morpheme-based lexical chunking in a recent study [9]. However, not until recent years have machine learning techniques been successfully applied to this particular task [49,31]. The rapid growth of large scale segmented corpora has brought in essential training data to enable the application of these techniques and the Bakeoffs,[2] in particular, have provided an international forum for examining and comparing their effectiveness in a comprehensive way [37,5,24], in addition to providing indispensable benchmark data sets for verifying and comparing the effectiveness of different supervised and unsupervised learning models.

Conventionally, supervised segmentation assumes a pre-segmented (or labeled) corpus for training a statistical model that can infer the optimal segmentation for an input sentence. Minimally, it assumes a bare vocabulary as a set of known words, either pre-defined or extracted from the pre-segmented training corpus. These words form the structural backbone of the model, and the optimal parameters of the model are to be obtained via training. However, regardless of whether the training is conducted on labeled or unlabeled data, or even both, its purpose is to determine the parameters in association with word sequences, not to determine the words. In principle, this kind of supervised training for a word-based model can hardly bring in any intrinsic mechanism for inferring OOV words. Unfortunately, among the two main causes of segmentation error, OOV words account for several times more errors than ambiguities [15], because a word-based statistical model trained on a labeled corpus of only known words has a strong power of disambiguation via probability estimation but little means to deal with unknown words.

In contrast, unsupervised segmentation does not rely on any given language resource such as a pre-defined vocabulary or a pre-segmented corpus. It is assumed to perform without any labeled data for training. In fact, it is intended to derive a vocabulary from scratch from unsegmented texts, in a way to estimate the likelihood of a substring being a true word by virtue of some pre-defined heuristics or a goodness measure. Usually, a goodness measure for this purpose is formulated in terms of statistical theory to capture our empirical observations of language characteristics in real data, rather than to express theoretical linguistic insights. Thus, it is not a surprise that linguistically-driven heuristic rules are often applied to remedy some idiosyncratic shortcomings of a statistics-based goodness measure for performance enhancement. Furthermore, it is worth noting that unsupervised segmentation is different from, and more complex than, word extraction. The former aims to carry out the segmentation task for a text, for which a decoding algorithm is indispensable; whereas the latter only needs to derive as the final output a list of word candidates from an unsegmented corpus [3,50,6], and hence may or may not involve segmenting the whole corpus into individual words.

Several studies have explored unsupervised segmentation of Chinese texts into words by various means and for various purposes [38,34,10,8,33,40,19]. These studies were formulated in very diverse ways, involving many kinds of heuristic rules. To our knowledge, however, there has not yet been any comprehensive evaluation to examine and compare the performance of different approaches in a consistent way using authoritative large scale "gold standard" data sets, such as the multi-standard ones for the Bakeoffs. Certainly, it is also more than interesting to take a close look at how these approaches correspond with different segmentation standards.

Considering that a statistical goodness measure represents, to a great extent, human observations of the global distributional characteristics of substrings throughout a given corpus of raw texts, one may proceed to exploit such characteristics to facilitate supervised learning of word segmentation. So far, supervised learning has only made use of local information about individual characters and/or substrings within the scope of a sentence, resorting to little global information derived from a whole large scale corpus.

This study explores the role that such global information, derived by various goodness measures, can play in both unsupervised and supervised segmentation. In particular, it focuses on examining four representative goodness measures for word discovery, namely, frequency [27], description length gain [21,20], accessor variety [6,7], and boundary entropy [44,3,16,19], aiming at exploring an effective way of applying them to enhance supervised segmentation.

Each of these measures is integrated into two baseline segmentation frameworks for a comprehensive evaluation. One is a generalized decoding algorithm to realize unsupervised segmentation with a goodness measure as objective function. The other is the conditional random fields (CRFs) model [23] for supervised segmentation via character tagging, conventionally trained only on a pre-segmented corpus. The latter is a state-of-the-art approach that has set new performance records in the field, as illustrated in [52,55], although its efficiency is yet to be further enhanced by various means [58,59]. All scores given by the goodness measures are discretized in the same way for use as feature values in the CRFs model. No other heuristic rules or prior knowledge are involved in this framework, so as to ensure a fair way of comparing the effectiveness of these goodness measures in enhancing the performance of the CRFs model. In this situation, among all evaluation measures in use, the error reduction rates indicating any further improvement over the existing performance records are particularly worth highlighting. All evaluations, for both supervised and unsupervised segmentation, are conducted on the benchmark data sets for Bakeoff-3 [24]. The main reason is that they are significantly larger than others and hence can provide technically more reliable evaluation results.

[2] The International Chinese Word Segmentation Bakeoffs, at http://www.sighan.org/{bakeoff2003,bakeoff2005,bakeoff2006} and http://www.chinalanguage.gov.cn/bakeoff 08/bakeoff-08_basic.html, conventionally referred to as Bakeoff -1, -2, -3 and -4, respectively.