Expert Systems with Applications 37 (2010) 3553–3560

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Chinese text segmentation: A hybrid approach using transductive learning and statistical association measures

Richard Tzong-Han Tsai

Department of Computer Science and Engineering, Yuan Ze University, Taiwan

ARTICLE INFO

Keywords:
Chinese word segmentation
Unlabeled data
Transductive learning
Association measure

ABSTRACT

Chinese text segmentation (CTS) is a fundamental step in building any Chinese or cross-language information retrieval system. This paper identifies and proposes solutions to two main challenges facing today's CTS systems: segmenting words longer than the context window and identifying words not derived from affixation or composition. Our methods exploit unlabeled data, making them scalable at little extra cost. To tackle the first problem, we use a transductive learning approach to automatically construct a dictionary, and then refine it by improving its test set coverage while reducing its over-fitting tendency. In addition, we incorporate frequency information to discriminate overlapping matching words. For the second problem, we employ statistical association measures non-parametrically through a natural but novel feature representation scheme. To demonstrate the generality of our approach, we verify our system on the most reputable CTS evaluation standard – the SIGHAN bakeoff, which contains datasets in both traditional and simplified Chinese. These datasets are provided by representative academic or industrial research institutes. The experimental results show that with only training data and unlabeled test data and with no external dictionaries, our approach effectively overcomes the above-mentioned problems and reduces segmentation errors by an average of 27.8% compared with the traditional approach. Notably, our approach improves the recall of new words, the most informative words, by 4.7% on average. Also, our approach outperforms the best SIGHAN CTS system, which requires many external resources. Additional analysis shows that our approach has the potential to gain accuracy as the test data increases.

1. Introduction

It is well known that Chinese sentences consist of continuous strings of characters with no delimiting spaces separating individual words. This makes it difficult for computers to process Chinese linguistic data unless it has been previously separated into individual words. Chinese text segmentation (CTS) is therefore a crucial first step before further linguistic analysis can begin, and automatic CTS is of particular interest (Kwok, 1999).

Approaches to Chinese text segmentation (CTS) fall roughly into two categories: heuristic dictionary-based and statistical machine learning methods. In dictionary-based methods, a predefined dictionary is used along with hand-generated rules for segmenting input sequences. Unfortunately, dictionary approaches are often limited by the impossibility of creating a lexicon that includes all possible Chinese words.

Thanks to the annotation efforts of many researchers in the computational linguistics field, several annotated corpora for CTS have been collected. In 2003, SIGHAN (Sproat & Emerson, 2003), a special interest group belonging to the Association for Computational Linguistics, held the first Chinese word segmentation bakeoff and released datasets in both traditional and simplified Chinese for training and evaluating CTS systems created by institutes specializing in Chinese language processing, these high-quality datasets are large enough to train machine learning models. Each has its own segmentation standard. Simultaneously, affordable off-the-shelf computer hardware has become powerful enough to execute machine learning programs consuming up to 32 GB of RAM. These two realities have made machine learning, which is more flexible and effective than dictionary-based approaches, a viable option for CTS today.

Machine learning can be divided into generative-model-based and discriminative-model-based types. Discriminative models such as maximum entropy models (MaxEnt) and conditional random fields (CRF) models provide a natural framework for incorporating arbitrary features. These models formulate the Chinese text segmentation problem as a character tagging problem (Peng & Feng, 2004; Tseng, Chang, Andrew, Jurafsky, & Manning, 2005; Xue & Shen, 2003). That is, assigning each character a tag to indicate its position in a word. The simplest formulation has two tags: beginning and inside (not beginning), denoted as B and I, respectively.

E-mail address: thtsai@saturn.yzu.edu.tw

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.10.004

Most statistical models use *n*-gram features to represent the current character and its context by describing the relative positions of certain characters to the current character within a context window. Note that the *n*-gram here is a little different from the one used in language models. The latter refers to the Markov order of the model plus one. Because of *n*-gram features, MaxEnt and CRF models are able to identify most part of out-of-vocabulary (OOV) words.

In spite of their popularity, *n*-gram features have some weaknesses. Short *n*-gram features may cause CTS systems to over-segment words longer than *n* (referred as longer words thereafter), especially when processing a long compound word. With a limited window size, it is hard to discriminate whether an *n*-gram should be segmented as a word or part of a longer word, especially when both examples occur in the training corpus. However, employing longer *n*-gram features demands a large amount of system memory and may suffer from data sparseness.

Another weakness of *n*-gram features is their poor ability to recognize new words which are the smallest independent units of meaning, such as names of newly famous people, movies, products, etc. According to Khoo and Loh (2002), these words are denoted as simple words. *N*-gram features are more effective at detecting new compound words rather than new simple words because simple words often do not contain known affixes or components, which are the main clues *n*-gram features rely on to detect new words.

In this paper we propose effective strategies to deal with the two above-mentioned weaknesses of *n*-gram features. One is based on transductive learning techniques and the other is based on association measures. According to our experimental results on the SIGHAN datasets, our system outperforms all other top SIGHAN systems, effectively overcoming the deficiencies of *n*-gram features.

2. Previous work

In this section we describe previous attempts to compensate for the two drawbacks of using *n*-gram features, namely, over-segmentation of longer words and poor ability to detect new simple words.

2.1. Segmenting longer words with dictionary features

Using dictionary features is a straightforward way to enhance longer word segmentation. These features indicate whether the context matches a word in a given dictionary. The simplest way to create a dictionary is to collect all words in the training set. This is referred to as a *training dictionary*. Since the training dictionary has lower coverage of words in the test set, solely using the training dictionary results in a large number of OOV words, which are often incorrectly segmented into single-character or overly-short words (Chen & Bai, 1998). This is referred to as the *over-segmentation* problem.

To overcome this problem, dictionary coverage of the test set must also be improved. One possible course to achieve this is to add an external dictionary whose content is independent of the training and test corpora. However, such dictionaries are very costly to maintain and not applicable if they follow different segmentation standards from that of the corpus.

Another more economical approach is to apply new word identification (NWI), which extracts new words from the test set and adds them to the dictionary. We denote the final product of this process as a *transductive dictionary* since this procedure follows the general concepts of transductive learning. Currently there are two NWI approaches: one is independent of CTS, whereas the other is part of it. Gao et al. (2004) are one representative example of the first type. They used a linear model which combined different source-channel models, including an NWI model.

Peng and Feng (2004) used the second type of NWI. They used their baseline model to tag the test set, segmented the results, and added new words with confidence over 0.9 to the dictionary.

Another problem that results from using dictionary information is the word overlap problem (Chen & Bai, 1998), which refers to a phrase "cde", where both "cd" and "de" are in the dictionary. Chen and Bai (1998) solved this conflict by selecting the word with the higher (frequency $\times$ length). We utilize this idea to enhance the dictionary features in our machine-learning-based (ML-based) CTS models.

In this paper, to reduce the dictionary coverage difference between the training and test sets, we create a balanced transductive dictionary that identifies new words in the test set. In addition, we also develop several useful dictionary features with frequency information and demonstrate their effectiveness.

2.2. Detecting new simple words with association measures

New simple words are more likely to be missed by the baseline *n*-gram model because their morphemes' morphological tendencies are not always consistent with those recorded by *n*-gram features. For instance, supposing 天平 tian-ping (libra) only appears as individual words in the training set, the baseline model will not be able to identify 熊天平 xiong-tian-ping (a singer's name) because 熊天平 is a simple word and the morphological tendency of 天平 (be a suffix of 天平) is not consistent with that in the model (be a individual word).

In the English language, there is a similar phenomenon called multi-word expressions (MWEs). Choueka (1988) described MWE as connected collocations: a sequence of neighboring words "whose exact and unambiguous meaning or connotation cannot be derived from the meaning or connotation of its components". This suggests some resemblance between MWEs and simple words. Some pioneers have proposed MWE identification methods based on association measures (AM), such as pointwise mutual information (Church & Hanks, 1990), *c*-values (Frantzi, Ananiadou, & Mima, 2000), or likelihood ratios (Dunning, 1993). We adopt a method based on these metrics to recognize simple Chinese words.

AM-based NWI methods can be divided into two types – filtering and merging. The former uses association measures to measure the likelihood that a word candidate can be divided into two individual words and to filter out candidates with scores lower than a threshold. The latter merges character segments in a bottom-up fashion. Candidates with the highest association measures are merged in order. Examples of filtering and merging methods are Peng and Schuurmans (2001) and Ma and Chen (2003), respectively.

However, these two methods suffer from several shared drawbacks: in order for association measures to be applied to ranking character segment pairs, the metrics must be normalized because they are calculated from character segment pairs of various lengths. Normalization is based on heuristics, which have no theoretical support. In addition, these NWI methods ignore the relationships among *n*-grams or other contextual information and labels (such as B and I), which are usually abundant in annotated corpora. Instead, they only use annotation data to select thresholds or create merging rules.

In this paper, we propose a fix for these shortcomings of previous AM-based methods using a CRF model. This idea came from our observation that using association measures to detect breaks in character segments is analogous to using character-based tagging to detect whether the next character's tag is B. This led us to conclude that Association Measures could be formulated as features in a CRF model.