Information Processing and Management 49 (2013) 626–636

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

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

Probabilistic Chinese word segmentation with non-local information and stochastic training

Xu Sun [a,b,*], Yaozhong Zhang [c], Takuya Matsuzaki [d], Yoshimasa Tsuruoka [e], Jun'ichi Tsujii [f]

[a] Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, Beijing, China
[b] School of EECS, Peking University, Beijing, China
[c] Dept of Computer Science, The University of Tokyo, Tokyo, Japan
[d] National Institute of Informatics, Tokyo, Japan
[e] Dept of EEIS, The University of Tokyo, Tokyo, Japan
[f] Microsoft Research Asia, Haidian District, Beijing, China

ARTICLE INFO

Article history:
Received 29 May 2011
Received in revised form 15 October 2012
Accepted 10 December 2012

Keywords:
Word segmentation
Natural language processing
Conditional random fields
Latent conditional random fields
Online training

ABSTRACT

In this article, we focus on Chinese word segmentation by systematically incorporating non-local information based on latent variables and word-level features. Differing from previous work which captures non-local information by using semi-Markov models, we propose an alternative method for modeling non-local information: a latent variable word segmenter employing word-level features. In order to reduce computational complexity of learning non-local information, we further present an improved online training method, which can arrive the same objective optimum with a significantly accelerated training speed. We find that the proposed method can help the learning of long range dependencies and improve the segmentation quality of long words (for example, complicated named entities). Experimental results demonstrate that the proposed method is effective. With this improvement, evaluations on the data of the second SIGHAN CWS bakeoff show that our system is competitive with the state-of-the-art systems.

1. Introduction

In most natural language processing tasks, words are the basic units to process. Since Chinese sentences are written as continuous sequences of characters, segmenting a character sequence into a word sequence is the first step for most Chinese processing applications. In this paper, we study the problem of Chinese word segmentation (CWS), which aims to find these basic units (words[1]) for a given sentence in Chinese.

1.1. Ambiguities and long words

Chinese character sequences are often ambiguous in their segmentation, and out-of-vocabulary (OOV) words are a major source of the ambiguity. Typical examples of OOV words include named entities (e.g., organization names, person names, and location names). Those named entities may be very long, and a difficult case occurs when a long word W ($|W| \geqslant 4$) consists of some words which can be separate words on their own; in such cases an automatic segmenter may split the OOV word into individual words. Named entities are hard to segment. First, named entities are typically long character strings. In some criteria of word segmentation (e.g., MSR standard), a long named entity is supposed to be a single word, and in this

* Corresponding author at: Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, Beijing, China.
E-mail address: xusun@pku.edu.cn (X. Sun).

[1] Following previous work, in this paper, *words* can also refer to multi-word expressions, including proper names, long named entities, idioms, etc.

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2012.12.003

Computer Committee of International Federation of Automatic Control
国际自动控制联合会计算机委员会

Fig. 1. An example of a complex single "word".

setting we should not segment a long named entity into small segments. A typical error in word segmentation is that the long character strings of named entities are incorrectly segmented into small segments. Second, named entities may consist of many lexicon words, which are already registered in dictionaries. Such lexicon words can frequently appear in other sentences, not necessarily with named entities. Hence, it is probable that a named entity will be incorrectly segmented into multiple lexicon words.

For example, the word shown in Fig. 1 is one of the organization names in the Microsoft Research corpus. Its length is 13 and it contains more than six individual words, but it should be treated as a single word. Proper recognition of long OOV words is important not only for word segmentation, but also for a variety of other purposes, e.g., full-text indexing. However, as is illustrated, recognizing long words (without sacrificing the performance on short words) is challenging.

1.2. Existing methods

Conventional approaches to Chinese word segmentation treat the problem as a character-based labeling task (Xue, 2003; Zhao, Huang, Li, & Lu, 2010; Zhao & Kit, 2011). Labels are assigned to each character in the sentence, indicating whether the character x_i is the start ($Label_i = B$), middle or end of a multi-character word ($Label_i = C$). A popular discriminative model that has been used for this task is the conditional random fields (CRFs) (Lafferty, McCallum, & Pereira, 2001), starting with the work of Peng, Feng, and McCallum (2004). In the Second International Chinese Word Segmentation Bakeoff (the second SIGHAN CWS bakeoff) (Emerson, 2005), two of the highest scoring systems in the closed track competition were based on a CRF model (Asahara et al., 2005; Tseng, Chang, & Andrew, 2005).

While the CRF model is quite effective compared with other models designed for CWS, it may be limited by its restrictive independence assumptions on non-adjacent labels. Although the window can in principle be widened by increasing the Markov order, this may not be a practical solution, because the complexity of training and decoding a linear-chain CRF grows exponentially with the Markov order (Andrew, 2006).

To address this difficulty, a choice is to relax the Markov assumption by using the semi-Markov conditional random field model (semi-CRF) (Sarawagi & Cohen, 2004). Despite the theoretical advantage of semi-CRFs over CRFs, however, some previous studies (Andrew, 2006; Liang, 2005) exploring the use of semi-CRFs for Chinese word segmentation did not find significant gains over the standard CRF models. As discussed in Andrew (2006), the reason may be that despite the greater representational power of the semi-CRF, there are some valuable features that can be more naturally expressed in a character-based labeling model. For example, on a CRF model, one might use the feature "the current character x_i is X and the current label $Label_i$ is C". This feature may be helpful in CWS for generalizing to new words. This type of features is less natural in a semi-CRF, since in that case local features $\varphi(y_i, y_{i+1}, x)$ are defined on pairs of adjacent words.

1.3. Proposals

In this paper, instead of using semi-Markov models, we describe an alternative based on latent variables and word-level features to learn long range dependencies in Chinese word segmentation. We use the latent conditional random fields (LDCRFs) (Morency, Quattoni, & Darrell, 2007; Petrov & Klein, 2008), which use latent variables to carry additional information that may not be expressed by those original labels, and therefore try to build more complicated or longer dependencies. This is especially meaningful in CWS, because the used labels are quite few: $Label(y) \in \{B, C\}$, where B signifies *beginning a word* and C signifies *the continuation of a word*.[2] For example, by using LDCRFs, the aforementioned feature may turn to "the current character x_i is X, $Label_i = C$, and $LatentVariable_i = LV$". The current latent variable LV may strongly depend on the previous one or many latent variables, and therefore the system can model the long range dependencies which may not be captured by those very simple labels.

Further, we use word-level features to model non-local information. Compared with the traditional character-based features, the word-based features can learn non-local information better. The word-based features are extracted from the training data only, and there is no need to use extra resources like lexicons. Since character and word information have their different advantages in word segmentation, we use both character-based features and word-based features.

Not surprisingly, learning non-local information increases computational complexity in the training. The use of latent variables can increase the inference lattice and slow down the training speed. The use of non-local features can increase the feature dimension and further slow down the training speed. Traditional batch training methods are very slow on training LDCRFs (Petrov & Klein, 2008; Sun, Matsuzaki, Okanohara, & Tsujii, 2009a). For example, Petrov and Klein (2008) mentioned both time and memory cost problems on training LDCRFs for natural language parsing. Also, Sun et al. (2009a) showed that the training of LDCRFs is computationally expensive on large scale problems. To accelerate training speed, we further present an improved online training method for optimizing LDCRFs. We demonstrate that the improved online training

[2] In practice, one may add a few extra labels based on linguistic intuitions (Xue, 2003).