Artificial Intelligence 230 (2016) 173–191

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Exploiting meta features for dependency parsing and part-of-speech tagging

Wenliang Chen [a], Min Zhang [a,*], Yue Zhang [b], Xiangyu Duan [a]

[a] School of Computer Science and Technology, Soochow University, China
[b] Singapore University of Technology and Design, Singapore

ARTICLE INFO

Article history:
Received 31 October 2013
Received in revised form 8 August 2015
Accepted 7 September 2015
Available online 27 October 2015

Keywords:
Dependency parsing
Natural language processing
Meta-features
Part-of-speech tagging
Semi-supervised approach

ABSTRACT

In recent years, discriminative methods have achieved much progress in natural language processing tasks, such as parsing, part-of-speech tagging, and word segmentation. For these methods, conventional features in a relatively high dimensional feature space may suffer from sparseness and thus exhibit less discriminative power on unseen data. This article presents a learning framework of feature transformation, addressing the sparseness problem by transforming sparse conventional base features into less sparse high-level features (i.e. meta features) with the help of a large amount of automatically annotated data. The meta features are derived by bucketing similar base features according to the frequency in large data, and used together with base features in our final system. We apply the framework to part-of-speech tagging and dependency parsing. Experimental results show that our systems perform better than the baseline systems in both tasks on standard evaluation. For the dependency parsing task, our parsers achieve state-of-the-art accuracy on the Chinese data and comparable accuracy with the best known systems on the English data. Further analysis indicates that our proposed approach is effective in processing unseen data and features.

1. Introduction

Discriminative methods have been highly successful in solving structured prediction tasks in natural language processing (NLP), such as parsing, part-of-speech tagging, and word segmentation [1–10]. An important reason is that discriminative models accommodate rich features without constraints such as probabilistic independence assumptions between features. Taking dependency parsing as an example, recent advances in parsing accuracies have been driven by the incorporation of rich non-local features in discriminative models [4–6].

One drawback of rich and complex features, however, is that they can be sparse and rare in unseen test data. For example, a lexicalized feature, which contains a specific word in the training data, cannot be instantiated on out-of-vocabulary words in test data. Given a limited amount of manually annotated training data, it is even harder for the more complex bilexical or trilexical features, which contain more than one word, to be fully utilized in unknown test data.

Several methods have been proposed to address the sparseness issue by leveraging large-scale unannotated data. In particular, word clusters [11,12] have been used as additional features in discriminative models to alleviate the sparseness of lexicalized features, leading to improved accuracies in named entity recognition [13] and dependency parsing [14]. Recently, word embeddings [15] have also been used as less-sparse word features to improve part-of-speech tagging [16,17], named

* Corresponding author.
E-mail addresses: wlchen@suda.edu.cn (W. Chen), minzhang@suda.edu.cn (M. Zhang).

http://dx.doi.org/10.1016/j.artint.2015.09.002
0004-3702/

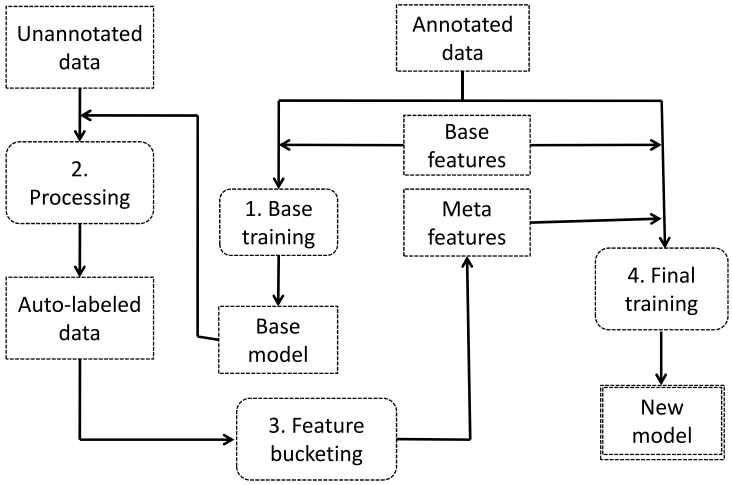

Fig. 1. Learning framework of feature transformation.

entity recognition [18,19], and dependency parsing [20], leading to similar improvements compared with Brown clusters. Obtained from large-scale data, word clusters and embeddings effectively enlarge the vocabulary, allowing words out of the training data to be better handled at test time.

There has been work on reducing the sparseness of structures beyond words. In particular, Chen et al. [21] extract subtree structures from auto-parsed dependency trees, groups the subtrees into clusters by their frequencies and incorporates the resulting clusters into a dependency parser as less-sparse subtree features. The method leads to accuracy improvements that are comparable to the use of Brown word clusters. The same idea has been applied to word segmentation and pos-tagging by making less-sparse representations of word ngrams from automatically processed data [22]. Experiments show that the simple method gives significant improvements in accuracies.

We extend the idea above from task-specific structures to arbitrary features, obtaining less-sparse features by clustering feature instances from automatically annotated data. Specifically, the feature instances under each feature template are bucketed by their frequencies, and each bucket is taken as a cluster. The clusters contain information on structured feature templates, and statistical distributions of the structures over large automatically annotated data. They are used to form a set of less-sparse features, which we call **meta features**. Compared with task-specific clusters such as subtree features, meta features have two main advantages. First, they are *general* and task independent. As a result, the method of this article can be used for any structured prediction tasks. Second, they are relatively more comprehensive by covering all the feature templates in a discriminative model, which are designed to include the most important structured patterns for a task.

We apply meta features to two typical structured prediction tasks in NLP, namely dependency parsing and part-of-speech (POS) tagging. For the dependency parsing task, our method significantly outperforms the method of [21], achieving accuracies comparable to the best reported in the literature. For the POS tagging task, our method also perform better than strong baselines.

This article is a significant extension of a conference version [23], which focuses on the dependency parsing task. We reformulate the method as a general framework for structured prediction, and demonstrate the effectiveness by evaluation on POS tagging in addition to dependency parsing. We give more details of the method, and in-depth analysis of the results.

The rest of this article is organized as follows. Section 2 gives a overview of the learning framework for discriminative structured prediction. In Section 3 shows the application of the proposed framework to dependency parsing. In Section 4 shows the application of the framework to part-of-speech tagging. Section 3.4 and 4.4 describe the experiment settings and reports the experimental results of dependency parsing and part-of-speech tagging respectively. Section 5 discusses related work. Finally, in Section 6 we summarize the proposed approach.

2. Learning framework of feature transformation

As shown in Fig. 1, the learning framework consists of four steps: 1) *Base training*: training a baseline system using annotated data and base features; 2) *Processing*: using the baseline system to annotate a large amount of raw sentences and obtain auto-labeled data; 3) *Feature bucketing*: performing feature transformation from base features to meta features; 4) *Final training*: training a new system with both the base and meta features.

The key step of the framework is *feature bucketing*, in which we use a transformation function to group the base feature instances from automatically annotated data into clusters, and define a set of meta features based on the clusters. The feature transformation can be treated as a clustering problem that groups the features with similar properties into the same cluster. Different clustering methods can be used for bucketing similar base features. We use frequency-based bucketing, putting the base features that have similar frequency levels in automatically annotated data into the same bucket, and