Engineering Applications of Artificial Intelligence 41 (2015) 166–174

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Unsupervised word sense induction using rival penalized competitive learning

Yanzhou Huang [a,b], Xiaodong Shi [a,b,*], Jinsong Su [c], Yidong Chen [a,b], Guimin Huang [d]

[a] Fujian Key Lab of the Brain-like Intelligent Systems, Xiamen University, Xiamen 361005, PR China
[b] Department of Cognitive Science, School of Information Science and Technology, Xiamen University, Xiamen 361005, PR China
[c] School of Software, Xiamen University, Xiamen 361005, PR China
[d] Research Center on Data Science and Social Computing, Guilin University of Electronic Technology, Guilin 541004, PR China

ARTICLE INFO

Article history:
Received 2 April 2014
Received in revised form
21 December 2014
Accepted 3 February 2015

Keywords:
Natural language processing
Word sense induction
Multi-granularity semantic representation
Competitive learning

ABSTRACT

Word sense induction (WSI) aims to automatically identify different senses of an ambiguous word from its contexts. It is a nontrivial task to perform WSI in natural language processing because word sense ambiguity is pervasive in linguistic expressions. In this paper, we construct multi-granularity semantic spaces to learn the representations of ambiguous instances, in order to capture richer semantic knowledge during context modeling. In particular, we not only consider the semantic space of words, but the semantic space of word clusters and topics as well. Moreover, to circumvent the difficulty of selecting the number of word senses, we adapt a rival penalized competitive learning method to determine the number of word senses automatically via gradually repelling the redundant sense clusters. We validate the effectiveness of our method on several public WSI datasets and the results show that our method is able to improve the quality of WSI over several competitive baselines.

1. Introduction

Word sense induction (WSI) is crucial for many natural language processing (NLP) tasks as word sense ambiguity is prevalent in all natural languages. WSI and word sense disambiguation (WSD) are two related techniques for lexical semantic computation. The main distinction between the two techniques is that the former discriminates different senses without relying on a predefined sense inventory, while the latter assumes an ability to access an already known sense list. For discriminating different word senses, each occurrence of an ambiguous word is regarded as an ambiguous instance. WSI is to conduct unsupervised sense clustering among these ambiguous instances, and the number of the resulting clusters is explained as the number of induced word senses. We show an example of WSI of ambiguous word "ball" in Fig. 1.

We believe that WSI methods face two major challenges. First, the contextual semantic is not explored sufficiently when conducting context modeling. In general, shallow lexical features (e.g. unigrams or bigrams of words) surrounded the ambiguous instances that constitute an important ingredient in sense induction. However, such fine-grained semantic features will inevitably suffer from data sparsity problem. More advanced Bayesian methods use topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to learn topic distributions of ambiguous instances. Compared with the shallow features, topic features can capture latent topic structure and have more generalization ability in semantic representation. Topic models are able to exploit abstract conceptual structures; however, only using topic models may lose certain amount of unique lexical semantics during context modeling. Based on this, we believe that using contextual features derived from multi-granularity semantic spaces can reflect various aspects of the semantic knowledge of the contexts.

Second, the sense number of ambiguous words cannot be determined appropriately. Many popular clustering methods such as *k-means* algorithm require the cluster number to be pre-assigned precisely. However, in many practical applications, it becomes impossible to know the exact cluster number in advance, such that these clustering algorithms often result in poor performance (Dehkordi et al., 2009). More recently, the non-parametric Bayesian method (Lau et al., 2012) uses Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006) to learn the number of word senses automatically. However, it tends to induce larger number of word sense when comparing to the gold standard per ambiguous word on SEMEVAL-2010 WSI dataset (Lau et al., 2012). Hence, exploring a word sense clustering algorithm to learn appropriate sense numbers for ambiguous words is also crucial for WSI task.

In this paper, we want to overcome the two challenges of WSI mentioned above. We propose a novel WSI framework that automatically induce word senses for ambiguous words over multi-granularity semantic spaces without relying a pre-assigned

* Correspondence to: 422 South Siming Road, Department of Cognitive Science, Xiamen University, Xiamen 361005, China. Tel.: +86 18959288068.
E-mail address: mandel@xmu.edu.cn (X. Shi).

http://dx.doi.org/10.1016/j.engappai.2015.02.004
0952-1976/

S1. He plays the ball.
S2. I would like to join into the ball tonight.
S3. The ball is running very fast.

Fig. 1. An example of word sense induction of ambiguous word "ball". Each occurrence of "ball" is underlined and regarded as an ambiguous instance. In this example, the senses of the instances in S1 and S3 are the same (highlighted with blue), totally different from the one in S2 (highlighted with green). We conduct word sense induction to identify the sense number of ambiguous word "ball" and assign the three instances to their corresponding senses, i.e. ideally {S1,S3} and {S2}. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

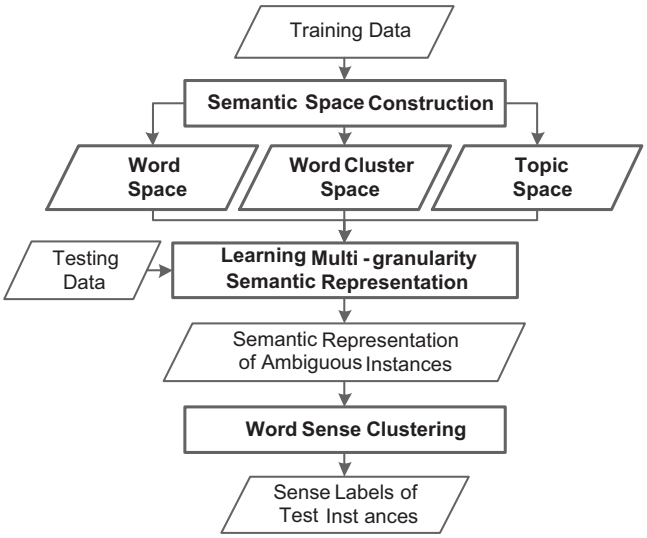

Fig. 2. The architecture of our proposed method in word sense induction.

number of word sense. In particular, our WSI framework runs in two steps: (1) learning multi-granularity semantic representations for ambiguous instances, and (2) context-based word sense clustering for ambiguous words.

For the first step, our main idea is that discriminating different word senses entails integrating diverse semantic granularities from the contexts. To be specific, we use Vector Space Model (Salton and Buckley, 1988) to learn the semantic representations of ambiguous instances, under the semantic space of words, word clusters and topics. Semantic distances among different semantic granularities are integrated in terms of a concatenation and a linear interpolation strategy (Section 3). For the second step, we adapt a rival penalized competitive learning (RPCL) method to determine the number of word senses automatically by gradually repelling the redundant sense clusters (Section 4). Once our algorithm matches a stopping condition, the centroid of the remaining clusters are considered as the representations of different word senses, and the number of remaining clusters are considered as the sense number induced for the ambiguous words. Fig. 2 summarizes the architecture of our proposed method for WSI.

Our method is able to improve the quality of WSI over several competitive baselines and the induced sense number is close to the gold standard sense. Especially, the main contributions of our work lie in two aspects, including (1) we integrate multi-granularity semantic spaces to represent the ambiguous instances without resorting to any external resources, and (2) instead of being pre-assigned a fixed number of word senses, our framework can automatically determine the sense number of ambiguous words.

The remainder of this paper is organized as follows: Section 2 summarizes and compares related work. Section 3 presents our method on how to learn a multi-granularity semantic spaces representation for each ambiguous instance. Section 4 elaborates the context-based word sense clustering for ambiguous words. Section 5 describes our experiments and shows results with discussions. Finally, Section 6 concludes and outlines future directions.

2. Related work

In this section, we give an overview of previous methods and the participating systems in the WSI task.

Overview of previous methods in WSI: In general, most of the researches in WSI are based on the Distributional Hypothesis (Harris, 1954), which indicates that words surrounded with similar contexts tend to have similar meanings. Previous methods have exploited various linguistic features such as first and second order context vectors (Purandare and Pedersen, 2004), bigrams and triplets of words (Purandare and Pedersen, 2004; Udani et al., 2005; Bordag, 2006), collocations (Klapaftis and Manandhar, 2008), and syntactic relations (Chen et al., 2009; Van de Cruys and Apidianaki, 2011) to conduct contexts modeling. To improve the usability of limited, narrow-domain corpora, Pinto et al. (2007) uses pointwise Mutual Information to construct a co-occurrence list to performing self-term expansion. Based on this contextual features, vector-based (Salton and Buckley, 1988; Purandare and Pedersen, 2004; Pedersen, 2007, 2010; Niu et al., 2007; Pinto et al., 2007) and graph-based (Agirre and Soroa, 2007b; Klapaftis and Manandhar, 2008; Korkontzelos and Manandhar, 2010; Klapaftis and Manandhar, 2010) models are applied to WSI.

More advanced Bayesian methods have been explored in recent years as the methods can discover latent topic structures from contexts without involving feature engineering. Brody and Lapata (2009) uses parametric LDA (Blei et al., 2003) to WSI task. The contexts of ambiguous instances are regarded as pseudo documents and their induced topic distributions are considered as the sense distributions. Yao and Van Durme (2011) further use non-parametric HDP (Teh et al., 2006) to learn the sense distributions. The advantage of this method is that it can automatically learn the number of word senses for each ambiguous word, as compared to LDA which needs to be pre-assigned a topic number in advance. Experiment results show that the HDP model is superior to standard LDA model. Lau et al. (2012) also show improvement in supervised *F*-score after incorporating position features in the HDP model. Charniak (2013) extends the naive Bayes model based on the idea that the more closer a word to the target word, the more relevant this word will be in WSI.

Overview of participating systems in WSI: The evaluation campaigns of WSI have been conducted in SemEval-2007 (Agirre and Soroa, 2007a), SemEval-2010 (Manandhar et al., 2010) and SemEval-2013 (Navigli and Vannella, 2013). As to the participating systems in SemEval-2007, their methods mainly use the vector-based and graph-based models to conduct WSI of target words. I2R (Niu et al., 2007) is the best induction system (vector-based) in supervised evaluation which uses part-of-speech of neighboring words, unordered words and local collocations to capture contextual information.

Considering the systems in SemEval-2010, the highest ranked system in supervised evaluation is *UoY* (graph-based) (Korkontzelos and Manandhar, 2010) where single nouns and noun pairs are included as vertices in the graph. Each cluster is taken to represent one of the senses of the target word. Note that *KSU KDD* (Elshamy et al., 2010) introduces the topic model LDA (Blei et al., 2003) to infer the topic distribution of each test instance and the *k-means* algorithm is applied to sense clustering.

In view of systems in SemEval-2013, participating systems are applied to web search result clustering. The best performing systems are developed from those HDP teams (Navigli and Vannella, 2013) which take advantage of the topic model HDP (Teh et al., 2006) to