Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 22 (2008) 330–345
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Semi-supervised learning integrated with classifier combination for word sense disambiguation

Anh-Cuong Le [a,*], Akira Shimazu [a], Van-Nam Huynh [b], Le-Minh Nguyen [a]

[a] School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[b] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

Received 13 November 2006; received in revised form 20 October 2007; accepted 6 November 2007
Available online 23 November 2007

Abstract

Word sense disambiguation (WSD) is the problem of determining the right sense of a polysemous word in a certain context. This paper investigates the use of unlabeled data for WSD within a framework of semi-supervised learning, in which labeled data is iteratively extended from unlabeled data. Focusing on this approach, we first explicitly identify and analyze three problems inherently occurred piecemeal in the general bootstrapping algorithm; namely the imbalance of training data, the confidence of new labeled examples, and the final classifier generation; all of which will be considered integratedly within a common framework of bootstrapping. We then propose solutions for these problems with the help of classifier combination strategies. This results in several new variants of the general bootstrapping algorithm. Experiments conducted on the English lexical samples of Senseval-2 and Senseval-3 show that the proposed solutions are effective in comparison with previous studies, and significantly improve supervised WSD.

Keywords: Semi-supervised learning; Word sense disambiguation; Computational linguistics

1. Introduction and motivation

The automatic disambiguation of word senses has been an interest and concern since the 1950s. Roughly speaking, WSD involves the association of a given word in a text or discourse with a particular sense among numerous potential senses of that word. As mentioned in Ide and Véronis (1998), this is an "intermediate task" necessary in accomplishing most natural language processing tasks, such as message understanding and human-machine communication, machine translation, information retrieval, etc. Since its inception, many methods involving WSD have been developed in the literature (for a survey, see, e.g., Ide and Véronis, 1998).

So far, many supervised machine learning algorithms have been used for the task of WSD, including Naïve Bayes, decision tree, exemplar-based, support vector machines, maximum entropy models, etc. (see, for

* Corresponding author. Present address: College of Technology, Vietnam National University, Hanoi, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam. Tel.: +81761511757; fax: +81761511149.
E-mail address: huynh@jaist.ac.jp (A.-C. Le).

0885-2308/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2007.11.001

example, Lee and Ng, 2002). Due to the difficulty or the cost of obtaining labeled data, whilst unlabeled data is abundant and cheap to collect, recently several WSD studies have tried to utilize unlabeled data to boost the performance of supervised learning (e.g. Mihalcea, 2004a; Zheng et al., 2005; Su et al., 2004; Pham et al., 2005). The process of using both labeled and unlabeled data to build a classifier is called *semi-supervised learning*.

In the following, we first briefly introduce general approaches in semi-supervised learning for WSD. Then, we explicitly describe problems that may occur in the so-called *general bootstrapping algorithm*, which is used in this paper, and then provide some ideas for tackling these problems.

1.1. Semi-supervised learning approaches

As semi-supervised learning requires less human effort for preparing annotated labeled data and potentially gives higher accuracy, it is of great interest both in theory and in practice. Semi-supervised learning methods use unlabeled data to either modify or re-prioritize hypotheses obtained from labeled data alone. In our opinion, the methods in semi-supervised learning can be grouped into two approaches, as follows.

In the first approach, the learners try to optimize parameters of the classification model using both labeled and unlabeled data. Miller and Uyar (1997), and Nigam et al. (2000) used a generative model for the classifier and used Expectation Maximization to estimate the model's parameters trained on both labeled and unlabeled data. Joachims (1999) used transductive inference for support vector machines to optimize performance on a specific test set, while Blum and Chawla (2001) constructed a graph based on the whole examples and used a minimum cut on the graph to yield an optimal labelling for the unlabeled examples.

In the second approach, learners follow a strategy in which the initial labeled data is iteratively extended, and finally a larger set of labeled data is obtained and used to generate the final classifier. From the literature review, we observe that a common method for enlarging labeled data is to use the classifier trained on the current labeled dataset to detect labels for unlabeled examples. Among those new labeled examples, some highly accurate ones are selected and added to the current labeled dataset. This process is iteratively repeated until there is no unlabeled example left, or until the number of iterations reaches a pre-defined threshold. Two well-known methods based on this approach are self-training (Yarowsky, 1995) and co-training (Blum and Mitchell, 1998).

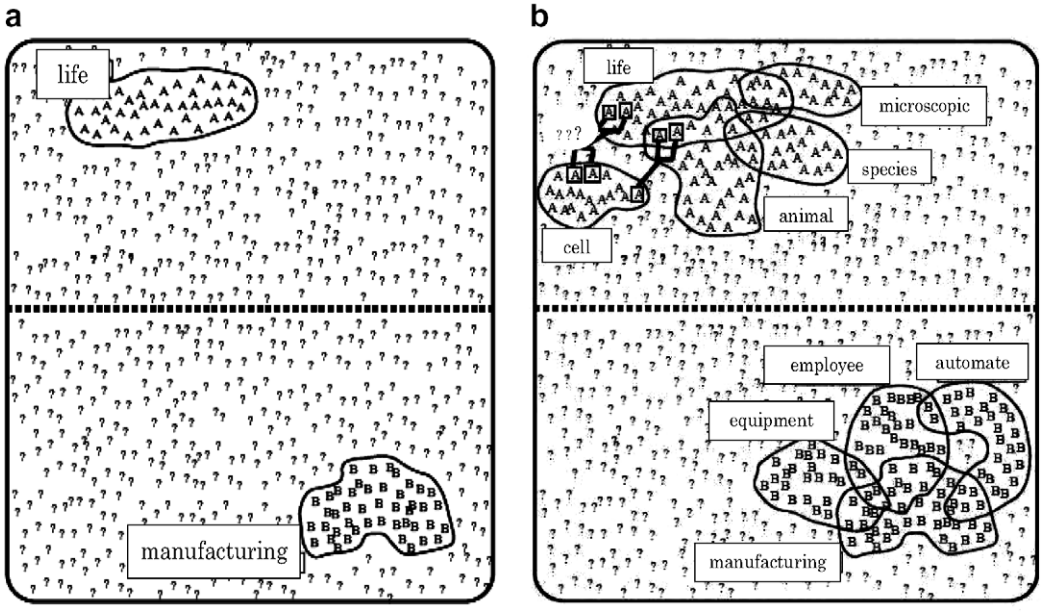

Fig. 1. A scheme to describe the process of iteratively extending labeled data.