Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 24 (2010) 461–473

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Adaptively entropy-based weighting classifiers in combination using Dempster–Shafer theory for word sense disambiguation ☆

Van-Nam Huynh [a,*], Tri Thanh Nguyen [b], Cuong Anh Le [b]

[a] *Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan*
[b] *College of Technology, Vietnam National University, 144 Xuan Thuy, Cau Giay District, Hanoi, Viet Nam*

Received 26 September 2008; received in revised form 27 November 2008; accepted 21 June 2009
Available online 27 June 2009

Abstract

In this paper we introduce an evidential reasoning based framework for weighted combination of classifiers for word sense disambiguation (WSD). Within this framework, we propose a new way of defining adaptively weights of individual classifiers based on ambiguity measures associated with their decisions with respect to each particular pattern under classification, where the ambiguity measure is defined by Shannon's entropy. We then apply the discounting-and-combination scheme in Dempster–Shafer theory of evidence to derive a consensus decision for the classification task at hand. Experimentally, we conduct two scenarios of combining classifiers with the discussed method of weighting. In the first scenario, each individual classifier corresponds to a well-known learning algorithm and all of them use the same representation of context regarding the target word to be disambiguated, while in the second scenario the same learning algorithm applied to individual classifiers but each of them uses a distinct representation of the target word. These experimental scenarios are tested on English lexical samples of Senseval-2 and Senseval-3 resulting in an improvement in overall accuracy.

Keywords: Computational linguistics; Classifier combination; Word sense disambiguation; Dempster's rule of combination; Entropy

1. Introduction

Polysemous words that have multiple senses or meanings appear pervasively in many natural languages. While it seems not much difficult for human beings to recognize the correct meaning of a polysemous word among its possible senses in a particular language given the context or discourse where the word occurs, the issue of automatic disambiguation of word senses is still one of the most challenging tasks in natural language processing (NLP) (Montoyo et al., 2005), though it has received much interest and concern from the research community

☆ This work was partially supported by a Grant-in-Aid for Scientific Research (No. 20500202) from the Japan Society of the Promotion of Science (JSPS) and FY-2008 JAIST International Joint Research Grant.

* Corresponding author. Tel.: +81 761511757.
E-mail address: huynh@jaist.ac.jp (V.-N. Huynh).

0885-2308/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2009.06.003

since the 1950s (see Ide and Véronis (1998) for an overview of WSD from then to the late 1990s). Roughly speaking, WSD is the task of associating a given word in a text or discourse with an appropriate sense among numerous possible senses of that word. This is only an "intermediate task" which necessarily accomplishes most NLP tasks such as grammatical analysis and lexicography in linguistic studies, or machine translation, man–machine communication, message understanding in language understanding applications (Ide and Véronis, 1998). Besides these directly language oriented applications, WSD also have potential uses in other applications involving knowledge engineering such as information retrieval, information extraction and text mining, and particularly is recently beginning to be applied in the topics of named-entity classification, co-reference determination, and acronym expansion (cf. Agirre and Edmonds, 2006; Bloehdorn and Andreas, 2004; Clough and Stevenson, 2004; Dill et al., 2003; Sanderson, 1994; Vossen et al., 2006).

So far, many approaches have been proposed for WSD in the literature. From a machine learning point of view, WSD is basically a classification problem and therefore it can directly benefit by the recent achievements from the machine learning community. As we have witnessed during the last two decades, many machine learning techniques and algorithms have been applied for WSD, including Naive Bayesian (NB) model, decision trees, exemplar-based model, support vector machines (SVM), maximum entropy models (MEM), etc. (Agirre and Edmonds, 2006; Lee and Ng, 2002; Leroy and Rindflesch, 2005; Mooney, 1996). On the other hand, as observed in studies of classification systems, the set of patterns misclassified by different learning algorithms or techniques would not necessarily overlap (Kittler et al., 1998). This means that different classifiers may potentially offer complementary information about patterns to be classified. In other words, features and classifiers of different types complement one another in classification performance. This observation highly motivated the interest in combining classifiers to build an ensemble classifier which would improve the performance of the individual classifiers. Particularly, classifier combination for WSD has been received considerable attention recently from the community as well (e.g. Escudero et al., 2000; Florian and Yarowsky, 2002; Hoste et al., 2002; Kilgarriff and Rosenzweig, 2000; Klein et al., 2002; Le et al., 2005; Le et al., 2007; Pedersen, 2000; Wang and Matsumoto, 2004).

Typically, there are two scenarios of combining classifiers mainly used in the literature (Kittler et al., 1998). The first approach is to use different learning algorithms for different classifiers operating on the same representation of the input pattern or on the same single data set, while the second approach aims to have all classifiers using a single learning algorithm but operating on different representations of the input pattern or different subsets of instances of the training data. In the context of WSD, the work by Klein et al. (2002), Florian and Yarowsky (2002), and Escudero et al. (2000) can be grouped into the first scenario. Whilst the studies given in Le et al. (2005), Le et al. (2007), Pedersen (2000) can be considered as belonging to the second scenario. Also, Wang and Matsumoto (2004) used similar sets of features as in Pedersen (2000) and proposed a new voting strategy based on kNN method.

In addition, an important research issue in combining classifiers is what combination strategy should be used to derive an ensemble classifier. In Kittler et al. (1998), the authors proposed a common theoretical framework for combining classifiers which leads to many commonly used decision rules used in practice. Their framework is essentially based on the Bayesian theory and well-known mathematical approximations which are appropriately used to obtain other decision rules from the two basic combination schemes. On the other hand, when the classifier outputs are interpreted as evidence or belief values for making the classification decision, Dempster's combination rule in the Dempster–Shafer theory of evidence (D–S theory, for short) offers a powerful tool for combining evidence from multiple sources of information for decision making (Al-Ani and Deriche, 2002; Bell et al., 2005; Denoeux, 1995; Denoeux, 2000; Le et al., 2007; Rogova, 1994; Xu et al., 1992). Despite the differences in approach and interpretation, almost D–S theory based methods of classifier combination assume the involved individual classifiers providing fully reliable sources of information for identifying the label of a particular input pattern. In other words, the issue of weighting individual classifiers in D–S theory based classifier combination has been ignored in previous studies. However, by observing that it is not always the case that all individual classifiers involved in a combination scenario completely agree on the classification decision, each of these classifiers does not by itself provide 100% certainty as the whole piece of evidence for identifying the label of the input pattern, therefore it should be weighted somehow before building a consensus decision. Fortunately, this weighting process can be modeled in D–S theory by the so-called discounting operator.