Information Processing and Management 45 (2009) 55–69

Contents lists available at ScienceDirect

Information Processing and Management

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

The bootstrapping of the Yarowsky algorithm in real corpora

Ricardo Sánchez-de-Madariaga *, José R. Fernández-del-Castillo

Departamento de Ciencias de la Computación, Escuela Técnica Superior de Ingeniería Informática, Politécnico, Campus Universitario, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain

ARTICLE INFO

Article history:
Received 14 January 2008
Received in revised form 29 May 2008
Accepted 16 July 2008
Available online 30 August 2008

Keywords:
Word sense disambiguation
Polysemy
Homograph
Knowledge acquisition bottleneck
Domain fluctuating corpora
Bootstrapping
Semi-supervised learning

ABSTRACT

The Yarowsky bootstrapping algorithm resolves the homograph-level word sense disambiguation (WSD) problem, which is the sense granularity level required for real natural language processing (NLP) applications. At the same time it resolves the knowledge acquisition bottleneck problem affecting most WSD algorithms and can be easily applied to foreign language corpora. However, this paper shows that the Yarowsky algorithm is significantly less accurate when applied to domain fluctuating, real corpora. This paper also introduces a new bootstrapping methodology that performs much better when applied to these corpora. The accuracy achieved in non-domain fluctuating corpora is not reached due to inherent domain fluctuation ambiguities.

1. Introduction

One specific problem in the field of computational linguistics is that of determining which particular sense of a word is being used in a given context. This problem is often referred to as *word sense disambiguation* (WSD). WSD is often viewed as a classification problem: the many occurrences of a given word form in some corpus are grouped into the classes represented by each of its possible senses. The set of possible senses of the word can be extracted from a generic dictionary, knowledge base or ontology; from an application-specific inventory; or it may even be a priori inexistent. Usually the context of each occurrence of the word in the corpus is used as the classifying criterion. WSD has been considered in many *natural language processing* (NLP) applications, both as a generic and as an integrated component. The most prominent include *machine translation* (MT), *information retrieval* (IR), *information extraction* (IE) *and text mining*, and *modern lexicography* (Agirre & Edmonds, 2006).

It is to be expected that in some highly constrained forms of texts, such as patents or some summaries, WSD will be of little or no relevance. However, in a wide range of less constrained texts WSD has a prominent role to play, particularly in important applications such as machine translation and for many kinds of information retrieval and information extraction.

There are three main approaches to the WSD problem. One of these approaches, referred to as *knowledge-based* or *dictionary-based*, relates to methods that use some kind of knowledge-rich resource such as a dictionary, thesaurus, concept hierarchy or lexical database. However, these sources of knowledge are in general independent of the target corpus being disambiguated. On the other hand, *supervised corpus-based* methods use (sense-tagged) corpus training material related in some way to the target corpus, instead of an independent external resource. Finally, *unsupervised corpus-based* methods cluster words in non-annotated raw corpora without any kind of training or independent source of knowledge evidence.

* Corresponding author. Tel.: +34 918856959/918856656; fax: +34 918856646.
E-mail addresses: ricardo.sanchez@uah.es (R. Sánchez-de-Madariaga), joseraul.castillo@uah.es (J.R. Fernández-del-Castillo).

0306-4573/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2008.07.002

While knowledge-based and supervised corpus-based methods categorize words based on some pre-existing sense inventory, unsupervised corpus-based methods do not. Frequently they are only told the total number of senses of a word form to be discriminated (Pedersen, 2006). Combinations exist of the knowledge-based approach with either of the other two, but not between these.

The supervised corpus-based (and also the knowledge-based) methods need hand-built resources like sense-tagged training corpora. These training data are difficult and expensive to produce. This problem is often referred to as the *knowledge acquisition bottleneck* (Màrquez, Escudero, Martínez, & Rigau, 2006) and it prevents many supervised methods from being applied to foreign language corpora, for example.

The number of senses distinguished by a WSD system is often called its granularity. While many systems, including those tested in the Senseval competition, use a fine-grained approach i.e. a high number of sense distinctions for each target word, it has been argued that systems used in real NLP applications only need a coarse-grained granularity with two sense distinctions (Ide & Wilks, 2006). The former is often called aspect polysemy and the latter is referred to as homograph-level or full polysemy. We will refer to these issues in more detail below.

The Yarowsky (1995) algorithm is a semi-supervised WSD method – i.e., it does not suffer the knowledge acquisition bottleneck and thus it can be easily ported to foreign unknown language corpora, or to corpora written in languages for which no specific resources are available, something very useful in a multi-lingual Web context. Moreover it resolves the homograph-level sense disambiguation to the accuracy level required by real NLP applications (above 95%). However, we will show that the Yarowsky algorithm is affected by the domain fluctuations present in real corpora, something that significantly reduces its accuracy in practice. We will also show that we can use a bootstrapping methodology to increase its accuracy in such real ambiguous corpora.

2. The Yarowsky bootstrapping algorithm

The Yarowsky (1995) algorithm uses a bootstrapping method for disambiguating homographs in non-annotated raw corpora. For this reason it is not strictly supervised, but neither is it an unsupervised method. Thus, it is often considered a semi-supervised algorithm. It uses a small *seed* set of labelled examples which are representative of each of the homograph senses.

2.1. Knowledge sources

The Yarowsky algorithm uses two different sources of linguistic knowledge. In relation to the set of knowledge sources for WSD listed in Agirre and Stevenson (2006) the Yarowsky algorithm uses one syntactic (*collocation*, KS 3) and one pragmatic/topical (*topical word association*, KS 10) source. In terms of the original Yarowsky (1995) paper it uses two "properties of human language": *one-sense-per collocation* and *one-sense-per discourse*.

The one-sense-per-collocation property states that word forms placed near the target word, called collocations, provide strong indication about its sense (Yarowsky, 1993). This effect varies depending on the type of collocation. It is strongest for immediately adjacent collocations (97%), and weakens with distance. It is much stronger for words in a predicate–argument relationship than for arbitrary associations at equivalent distance, and very much stronger for collocations with content words than for those with function words.

The one-sense-per-discourse property states that words show a strong tendency to exhibit only one-sense in any given document (Yarowsky, 1995), discourse (Gale, Church, & Yarowsky, 1992) or, we would say, domain. In Yarowsky (1995) a measure of this property is provided by means of its accuracy (if a word occurs more than once in a discourse, how often it takes on the majority sense for the discourse) and its applicability (how often the word occurs more than once in a discourse) for a set of 37,232 instances of 10 different homographs. Results show average 99.8% accuracy and 50.1% applicability. However, the Yarowsky algorithm uses one-sense-per-discourse in a flexible way; if there is probabilistic evidence regarding one-sense-per-collocation, it may be overridden.

2.2. The learning algorithm

The main idea behind the Yarowsky algorithm is to begin with a small set of correctly tagged seed examples representative of the two (or more) senses of a word and, using a combination of the one-sense-per-collocation and the one-sense-per-discourse tendencies in those examples, augment them with additional examples of each sense. After several iterations of this process (most) instances of the word in the original (untagged) corpus will be assigned their (probably correct) corresponding sense. Of course, the application of one-sense-per-collocation requires the availability of the context (represented by the $+/-k$-word window) of each instance of the target word in the original corpus.

There are several strategies for identifying the initial training seeds. A very simple one is hand-tagging a small subset of the instance contexts. Another (semi-automatic) procedure is to identify a small number (maybe only one for each sense) of word collocations representative of each sense and then to tag all instance contexts containing these collocates with the seed's sense label. An automatic procedure would be to extract collocate words from a dictionary's entry for the target sense. These words would occur with significantly greater frequency in the entry relative to the entire dictionary, and would therefore appear in the most reliable collocational relationships (Yarowsky, 1993) with the target word.