Information Processing and Management 51 (2015) 16–31

Contents lists available at ScienceDirect

Information Processing and Management

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

Word sense discrimination in information retrieval: A spectral clustering-based approach

Adrian-Gabriel Chifu [a,*], Florentina Hristea [b], Josiane Mothe [c], Marius Popescu [b]

[a] IRIT UMR5505, CNRS, Université de Toulouse, Université Paul Sabatier, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9, France
[b] University of Bucharest, Faculty of Mathematics and Computer Science, Department of Computer Science, Academiei 14, RO-010014 Bucharest, Romania
[c] IRIT UMR5505, CNRS, Université de Toulouse, Ecole Supérieure du Professorat et de l'Education, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9, France

ARTICLE INFO

Article history:
Received 29 July 2013
Received in revised form 7 August 2014
Accepted 22 October 2014
Available online 19 November 2014

Keywords:
Information retrieval
Word sense disambiguation
Word sense discrimination
Spectral clustering
High precision

ABSTRACT

Word sense ambiguity has been identified as a cause of poor precision in information retrieval (IR) systems. Word sense disambiguation and discrimination methods have been defined to help systems choose which documents should be retrieved in relation to an ambiguous query. However, the only approaches that show a genuine benefit for word sense discrimination or disambiguation in IR are generally supervised ones. In this paper we propose a new unsupervised method that uses word sense discrimination in IR. The method we develop is based on spectral clustering and reorders an initially retrieved document list by boosting documents that are semantically similar to the target query. For several TREC ad hoc collections we show that our method is useful in the case of queries which contain ambiguous terms. We are interested in improving the level of precision after 5, 10 and 30 retrieved documents (P@5, P@10, P@30) respectively. We show that precision can be improved by 8% above current state-of-the-art baselines. We also focus on poor performing queries.

1. Introduction

According to Lin (1997), "given a word, its context and its possible meanings, the problem of word sense disambiguation (WSD) is to determine the meaning of the word in that context".[1] Although WSD is generally easy for humans, it represents an issue for computers. The problem becomes even more difficult to solve when an ambiguous word occurs in short chunks of texts, such as a query in an information retrieval (IR) system.

Applying WSD to improve IR results is a well studied problem, but with controversial results as evidenced in the literature. Several authors have concluded that WSD in IR does not lead to significant retrieval performance improvement (Guyot, Falquet, Radhouani, & Benzineb, 2008; Sanderson, 1994). Various studies (Krovetz & Croft, 1992; Uzuner, Katz, & Yuret, 1999; Voorhees, 1993) have argued that the main problem in improving retrieval performance when using WSD is the inefficiency of the existing disambiguation algorithms, a problem which increases in the case of short queries.

In more recent years the issue remained "as to whether less than 90% accurate automated WSD can lead to improvements in retrieval effectiveness" (Stokoe, Oakes, & Tait, 2003). This remark refers primarily to the traditional task of WSD which

* Corresponding author.
E-mail addresses: adrian.chifu@irit.fr (A.-G. Chifu), fhristea@fmi.unibuc.ro (F. Hristea), josiane.mothe@irit.fr (J. Mothe), popescunmarius@gmail.com (M. Popescu).

[1] For a complete discussion of state-of-the-art WSD see the monograph (Agirre & Edmonds, 2006).

http://dx.doi.org/10.1016/j.ipm.2014.10.007
0306-4573/© 2014 Elsevier Ltd. All rights reserved.

identifies the meaning of the ambiguous word in context. This type of WSD is generally based on external sources, such as dictionaries or WordNet(WN)-like knowledge bases for labeling senses (Carpineto & Romano, 2012; Guyot et al., 2008) and is therefore knowledge-based.

Attempts to use knowledge-based WSD in IR have been numerous. In (Gonzalo, Verdejo, Chugur, & Cigarran, 1998) as well as in (Mihalcea & Moldovan, 2000) positive results were reported. These studies made use of semantic indexing based on WN synsets. However, they were all conducted on small data sets. As commented in (Ng, 2011), the evaluation is scaled up to a large test collection in (Stokoe et al., 2003) but the reported improvements are from a weak baseline. Positive results are also reported in (Kim, Seo, & Rim, 2004), although the quantum of improvements is small.

Zhong and Ng (2012) are among the few authors who more recently have expressed a growing belief in the benefits brought by WSD to IR – when using a supervised WSD technique. They constructed their supervised WSD system directly from parallel corpora. Experimental results on standard TREC collections show that, using the word senses tagged by this supervised WSD system, significant improvements over a state-of-the-art IR system can be obtained (Zhong & Ng, 2012). However, it is well known that supervised WSD cannot be used on a large scale in practice due to the absence of the necessary annotated/parallel corpora.

In contrast to all these authors, we are suggesting and investigating the usage of an unsupervised WSD technique. In this paper, we present an approach that aims at identifying clusters from similar contexts, where each cluster shows a polysemous word being used for a particular meaning. It is our belief that IR is an application for which this type of analysis is useful. Our approach is therefore not concerned with performing a straightforward WSD, but rather with differentiating among the meanings of an ambiguous word. Considering word sense discrimination rather than straightforward WSD avoids the use of external sources such as dictionaries or WN type synsets which are commonly used (Carpineto & Romano, 2012).

In this paper, we propose a new word sense discrimination method for IR based on spectral clustering. This state of the art clustering technique is now a hot topic; for example, Takacs and Demiris (2009) studied the use of spectral clustering in multi-agent systems while Borjigin and Guo (2012) recently discussed the cluster number determination in spectral clustering. Spectral clustering has been used in WSD for the first time by Popescu and Hristea (2011) who point out the importance of the clustering method used in unsupervised WSD.

We hereby show that WS discrimination based on spectral clustering outperforms the baseline when no WS discrimination is applied and also when using another unsupervised method (Naïve Bayes).

The present paper is organized as follows: in Section 2 we present the related works on WSD in IR; the focus is on unsupervised methods. Section 3 presents word sense discrimination based on spectral clustering. Section 4 presents the two step IR process using the proposed WS discrimination model. The evaluation is presented in Section 5. A more thorough analysis of the obtained results is performed in Section 6. Section 7 lays out the impact of automatically generated context on our proposed method. Section 8 concludes this paper.

2. Related work

Word sense ambiguity is a central concern in natural language processing (NLP). SENSEVAL defined the first evaluation framework for word sense disambiguation (WSD) in NLP (Kilgarriff, 1997). According to Kilgarriff and Rosenzweig (2000), SENSEVAL participants defined systems that can be classified into two categories: supervised systems, which use training instances of sense-tagged words and non-supervised systems. According to (Navigli, 2009), supervised systems are typically employed when a restricted number of words have to be disambiguated, while this type of system encounters more difficulties when all open-class words from a text have to be disambiguated. In addition to general WSD, many recent papers consider disambiguation of individuals (Artiles, Gonzalo, & Sekine, 2007; D'Angelo, Giuffrida, & Abramo, 2011; Piskorski, Wieloch, & Sydow, 2009) and disambiguation of place names (Leidner, 2007). Indeed, WSD has many applications, such as text processing, machine translation and information retrieval (IR), for which this type of disambiguation – proper names – can be useful (although not sufficient).

Krovetz and Croft (1992) were among the first to conduct a thorough analysis of ambiguity in IR. They used the CACM and TIME test collections and compared query word sense with word senses in retrieved documents. They found that sense mismatch occurs more often when the document is non-relevant to the query and when there are few common words bridging the query and the retrieved document. Another large scale study of word sense disambiguation in IR was conducted by Voorhees (1993). The automatic indexing process she developed used the "is-a" relations from WN and constructed vectors of senses to represent documents and queries. This approach was compared to a stem-based approach for 5 small collections (CACM, CISI, CRAN, MED, TIME). The results showed that the stem-based approach was superior overall, although the sense-based approach improved the results for some queries (Voorhees, 1993). Sanderson (1994) used the Reuters collection in his experiments and showed that disambiguation accuracy should be of at least 90% in order for it to be of practical use. He used pseudo-words in his experiments.

Schütze introduced word sense discrimination in IR (Schütze & Pedersen, 1995; Schütze, 1998). Moreover, Schütze considers that, in some cases, WSD can be defined as a two-stage process: first sense discrimination, then sense labeling. Sense discrimination aims at classifying the occurrences of a word into categories that share the same word sense. This type of approach is quite distinct from the traditional task of WSD, which, as already mentioned, classifies words relative to existing senses. Schütze and Pedersen (1995), Schütze (1998) created a lexical co-occurrence based thesaurus. They associated each