Data & Knowledge Engineering 90 (2014) 22–37

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Editorial

Wikimantic: Toward effective disambiguation and expansion of queries

Christopher Boston [a], Hui Fang [b], Sandra Carberry [a,*], Hao Wu [b], Xitong Liu [b]

[a] Department of Computer Science, University of Delaware, Newark, DE 19716, USA
[b] Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716, USA

ARTICLE INFO

Article history:
Received 5 June 2013
Accepted 1 July 2013
Available online 15 August 2013

Keywords:
Disambiguation
Query expansion
Search queries

ABSTRACT

This paper presents an implemented and evaluated methodology for disambiguating terms in search queries and for augmenting queries with expansion terms. By exploiting Wikipedia articles and their reference relations, our method is able to disambiguate terms in particularly short queries with few context words and to effectively expand queries for retrieval of short documents such as tweets. Our strategy can determine when a sequence of words should be treated as a single entity rather than as a sequence of individual entities. This work is part of a larger project to retrieve information graphics in response to user queries.

1. Introduction

The amount of information available electronically has increased dramatically over the past decade. Research efforts in information retrieval and information extraction have developed methods for identifying documents relevant to a user query and for extracting information from such documents. These efforts have focused on textual documents and, to some extent, on pictorial images, but information graphics (non-pictorial graphics such as bar charts and line graphs) have been largely ignored. This is particularly problematic in the case of information graphics in popular media such as newspapers and magazines since the content of the graphic is generally not repeated in the article text [6]. This is in contrast with scientific articles where the article's text explicitly refers to and explains its constituent graphics. Thus information graphics are an important knowledge source that should not be ignored.

Information graphics in popular media generally have a high-level message that they are intended to convey, such as that *Visa ranks first among credit cards in circulation*. We have developed a methodology for identifying this high-level message via a Bayesian network that reasons about the content of the graphic and the graphic's communicative signals, such as one bar being colored differently from the other bars in a bar chart. [10,38,5]. We are now developing a system for retrieving information graphics in response to a user query by relating the query to a combination of the graphic's intended message and any text in the graphic such as the graphic's caption, axis labels, etc.

To do this, we must first disambiguate the words in the query and expand the query with related words. Words have multiple senses; for example, the word *bank* can refer to a financial institution or to the side of a river, depending on the context in which it is used. Disambiguation is the problem of determining the correct sense in which a word is used in a sentence. Disambiguation requires a knowledge source and a fundamental issue in disambiguation is the choice of the knowledge source; these have ranged from structured resources such as WordNet to unlabelled corpora [28]. Another issue is the segmenting of a word sequence into units that should be considered as a single concept. For example, should *comic book* be treated as a single unit or as two separate words? And lastly, one has the issue of which approach to take, ranging from supervised approaches that learn from labeled training sets to unsupervised methods which use unlabelled corpora [28].

* Corresponding author at: Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA. Tel.: +1 302 831 1954.
E-mail address: carberry@cis.udel.edu (S. Carberry).

0169-023X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.datak.2013.07.004

A vexing problem for information retrieval is the *term mismatch* problem — namely that queries often do not contain the same words that are used to index a document. For example, a query might use the word *company* whereas the document uses the term *corporation*. Query expansion is the task of expanding a query with related words that make for a more robust query that will be more successfully matched with relevant documents. As with disambiguation, query expansion requires a knowledge source and the selection of an approach, which can vary from linguistic approaches such as following the links in WordNet to statistical approaches that compute similarity based on occurrences within the same document [8]. In addition, query expansion does not seek a single interpretation, as is done in disambiguation, but instead must rank possible expansion terms and select a top-rated set of terms for inclusion in the expanded query.

Although disambiguation and query expansion are common issues in text processing and information retrieval, our problem is exacerbated by the fact that even full sentence queries for information graphics are short and the textual content of the graphics (including words provided by the graphic's hypothesized intended message) is small compared with that of typical text documents. For example, although there are many existing methods that extract semantic information via disambiguation, most require large amounts of context terms or focus exclusively on named entities [4,11,27,31].

This article presents our research on Wikimantic, a system initially designed for disambiguation of short queries and which has been extended to query expansion. By a short query, we mean one with fewer than 15 words, of which typically a third are non-content words such as prepositions, articles, and auxiliary verbs. In the case of query expansion, we focus on the problem of microblog retrieval, whose goal is to retrieve relevant tweets for a given topic [29]. This search domain is chosen for several reasons. First, tweets are much shorter than traditional documents and are thus similar in length to the limited text (caption, axis labels, and intended message) that will be available for a system designed to retrieve information graphics. Second, it is a difficult search domain. The short length of tweets makes retrieval more susceptible to failure since relevant tweets may not contain exactly the same terms as those appearing in the query. Third, microblog retrieval is an important search domain. With the increasing usage of Twitter, users have urgent needs to make sense and make use of the information hidden in the tweets. However, since this is a new search problem, it has not yet been well-studied.

Our approach both to disambiguation and query expansion utilizes the huge information resource provided by Wikipedia, the online encyclopedia hosted by the Wikimedia Foundation. Wikipedia has garnered a lot of interest as a tool for facilitating disambiguation by providing a semantic web of hyperlinks and “disambiguation pages” that associate ambiguous terms with unambiguous articles [15,25,27,31]. For an encyclopedia, English Wikipedia is monolithic. It contains over 3.5 million articles which are connected by hundreds of millions of user-generated links. Although errors do exist in articles and link structure, Wikipedia's strong editing community does a good job of keeping them to a minimum. A study [37] of the revision histories in Wikipedia showed that although malicious vandalism does occur, particularly with political articles, it is quickly repaired by constant communal editing. In addition, the rapid editing of existing articles and the construction of new ones is a strength of Wikipedia and results in an up-to-date knowledge source.

Our work has several novel contributions which are important for information systems. First, we disambiguate text strings that to our knowledge are the shortest yet. Second, our method can determine when a sequence of words should be disambiguated as a single entity rather than as a sequence of individual disambiguations. Furthermore, our method does not rely on capitalization since users are notoriously poor at correct capitalization of terms in their queries; this is in contrast to the text of formal documents where correct capitalization can be used to identify sequences of words that represent a named entity. With respect to query expansion, our method produces good results for retrieval of short documents such as tweets and outperforms all of the systems in the Microblog Track 2011 at TREC [29]. Thus our system Wikimantic offers promise not only for our research on retrieval of information graphics but more generally for information systems that must semantically process short text.

2. Related work

Bunescu and Pasca are generally credited with being the first to use Wikipedia as a resource for disambiguation [4]. They formulated the disambiguation task to be a two step procedure where a system must (1) identify the salient terms in the text and (2) link them accurately. Though Bunescu and Pasca's work was initially limited to named entity disambiguation, Mihalcea later developed a more general system that linked all “interesting” terms [25].

Mihalcea's keyword extractor and disambiguator relied heavily on anchor text extracted from Wikipedia's inter-article links. When evaluating the disambiguator, Mihalcea gave it 85 random Wikipedia articles with the linked terms identified but the link data removed, and scored it based on its ability to guess the original target of each link. Mihalcea achieved an impressive F-measure of 87.73 [25], albeit with one caveat. Regenerating link targets is significantly easier than creating them from scratch, since the correct target must necessarily exist in Wikipedia and be particularly important to the context. Wikimantic is tasked with the more difficult problem of disambiguating all salient terms in the query indiscriminately.

Many Wikipedia based disambiguation systems use variants of Mihalcea's method which attempt to match terms in the text with anchor text from Wikipedia links [27,18,24]. When a match is found, the term is annotated with a copy of the link. Sometimes, a term will match anchor text from multiple conflicting links, in which case the system must choose among them. Milne and Witten's contribution was to look for terms that matched only non-conflicting links, and use those easy disambiguations to provide a better context for the more difficult ones [27]. Given a large text string, it's always possible to find at least one trivial term to start the process. However, short strings do not reliably contain trivial terms.

Ferragina and Scaiella [15] addressed this problem by employing a voting system that resolved all ambiguous terms simultaneously. They found that good results were attainable with text fragments as short as 30 words each. Although their results