Information Processing and Management 48 (2012) 755–766

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

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

Resolving ambiguity in biomedical text to improve summarization

Laura Plaza [a,*], Mark Stevenson [b], Alberto Díaz [a]

[a] Dpto. de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid, C/ Profesor José García Santesmases s/n, 28040 Madrid, Spain
[b] Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom

ARTICLE INFO

Article history:
Received 10 August 2010
Received in revised form 24 June 2011
Accepted 15 September 2011
Available online 29 November 2011

Keywords:
Biomedical summarization
Word sense disambiguation
WSD
Unified medical language system
UMLS
MetaMap

ABSTRACT

Access to the vast body of research literature that is now available on biomedicine and related fields can be improved with automatic summarization. This paper describes a summarization system for the biomedical domain that represents documents as graphs formed from concepts and relations in the UMLS Metathesaurus. This system has to deal with the ambiguities that occur in biomedical documents. We describe a variety of strategies that make use of MetaMap and Word Sense Disambiguation (WSD) to accurately map biomedical documents onto UMLS Metathesaurus concepts. Evaluation is carried out using a collection of 150 biomedical scientific articles from the BioMed Central corpus. We find that using WSD improves the quality of the summaries generated.

1. Introduction and background

A vast amount of literature on biomedicine and related fields is now available and growing at an increasing rate (Hunter & Cohen, 2006). Access to the information it contains is necessary for researchers and has also been shown to be useful for both health professionals and consumers (Lau & Coiera, 2008; Westbrook, Coiera, & Gosling, 2005). However, the amount of information available is now so large that tools are required in order to access it practically (Cohen & Hersh, 2005; Zweigenbaum, Demner-Fushman, Yu, & Cohen, 2007). Text summarization systems can improve this access (Hunter & Cohen, 2006; Reeve, Han, & Brooks, 2007). When no author's abstract is available, researchers can use summaries to determine whether a scientific article is of interest without having to read the entire document (Mani, 1999, 2001; Moens, 2000). Automatic summarization systems may be also used to assist scientists to write abstracts. Physicians can use summaries to identify treatment options, reducing the diagnosis time (Brooks & Sulimanoff, 2002). Reeve et al. (2007) states that there are two reasons for generating summaries from a full-text source, even when the author has created an abstract: (1) the abstract may not include relevant content from the full-text, and (2) there is no single "ideal" summary that meets the information needs of all users. Moreover, automatic summaries have been shown to improve indexing and categorization of biomedical literature, when used instead of the articles' abstracts (Gay, Kayaalp, & Aronson, 2005).

Summarization systems usually work with *text-level representations* of the document which consist of information that can be directly extracted from the document itself (Erkan & Radev, 2004; Mihalcea & Tarau, 2004). Studies have also demonstrated the benefit of richer *conceptual representations* (Fiszman, Rindflesch, & Kilicoglu, 2004; Plaza, Díaz, & Gervás, 2008), which represent documents using concepts instead of words. The representations may be enriched with semantic relations between the concepts (i.e. synonymy, hypernymy, homonymy or co-occurrence) to improve the quality of the summaries.

* Corresponding author. Tel.: +34 625638290; fax: +34 913947529.
E-mail addresses: lplazam@fdi.ucm.es (L. Plaza), m.stevenson@dcs.shef.ac.uk (M. Stevenson), albertodiaz@fdi.ucm.es (A. Díaz).

0306-4573/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2011.09.005

The Unified Medical Language System (UMLS) (Nelson, Powell, & Humphreys, 2002) has proved to be a useful knowledge source for summarization in the biomedical domain (Fiszman et al., 2004; Plaza et al., 2008; Reeve et al., 2007). When the UMLS is used, the vocabulary of the document being summarized has to be mapped onto the concepts it contains. This is made difficult by lexical ambiguity, the fact that words can have multiple meanings depending on the context in which they appear. Although it is often believed that technical domains contain less ambiguity than general ones (Farghlay & Hedin, 2003; Gale, Church, & Yarowsky, 1992), biomedical text has been shown to be highly ambiguous (Weeber, Mork, & Aronson, 2001). For example, the term "cold" is associated with several possible meanings in the UMLS Metathesaurus including 'common cold', 'cold sensation', 'cold temperature' and 'cold therapy'.

The majority of biomedical summarizers that employ the UMLS Metathesaurus use MetaMap (Aronson, 2001) to translate the text into UMLS concepts (Fiszman et al., 2004; Reeve et al., 2007) but do not attempt to resolve ambiguities when MetaMap returns multiple concepts. However, selecting the wrong meaning for ambiguous terms may affect the quality of the summaries generated.

This paper describes the application of various strategies for selecting UMLS concepts from the MetaMap output to improve a state-of-art biomedical summarization system. The summarizer (Plaza et al., 2008) is a graph-based method that uses the UMLS Metathesaurus to create conceptual representations. Strategies for selecting concepts from MetaMap include using Word Sense Disambiguation (WSD) (Agirre & Edmonds, 2006) to attempt to determine the meaning of words by examining their context. We find that using WSD improves the quality of the summaries generated.

The next section describes related work on summarization and WSD and also introduces the resources employed by the summarization and WSD systems used in this work. Section 3 describes our concept-based summarization algorithm. Section 4 presents the different WSD algorithms and strategies that have been tested to assign concepts from the UMLS. Section 5 describes the experimental environment of the study. Section 6 reports the results of the experiments and discusses these results. The final section provides concluding remarks and suggests future lines of work.

2. Related work

2.1. UMLS and MetaMap

The Unified Medical Language System (UMLS) (Nelson et al., 2002) is a collection of controlled vocabularies related to biomedicine and contains a wide range of information that can be used for Natural Language Processing (NLP). The UMLS comprises of three parts: the Specialist Lexicon, the Metathesaurus and the Semantic Network.

The **Specialist Lexicon** is a database of lexicographic information for use in NLP tasks that consists of a set of lexical entries with one entry for each spelling or set of spelling variants in a particular part of speech.

The **Metathesaurus** forms the backbone of the UMLS and is created by unifying over 100 controlled vocabularies and classification systems. It is organized around concepts, each of which represents a meaning and is assigned a Concept Unique Identifier (CUI). For example, the following CUIs are all associated with the term "cold": C0009443 'Common Cold', C0009264 'Cold Temperature' and C0234192 'Cold Sensation'.

The Metathesaurus comprises of several tables containing information about CUIs. These include the `MRREL` and `MRHIER` tables. The `MRREL` table lists relations between CUIs found in the various sources that are used to form the Metathesaurus. This table lists a range of different types of relations, including `child`, `parent`, `can be qualified by`, `related and possibly synonymous` and `other related`. For example, the `MRREL` table states that the concepts C0009443 'Common Cold' and C0027442 'Nasopharynx' are connected via the `other related` relation.

The `MRHIER` table in the Metathesaurus lists the hierarchies in which each CUI appears, and lists the entire path to the root of each hierarchy for the CUI.

The **Semantic Network** consists of a set of categories (or semantic types) that provides a consistent categorization of the concepts in the Metathesaurus, along with a set of relationships (or semantic relations) that exist between the semantic types. For example, the concept C0009443 'Common Cold' is classified in the semantic type 'Disease or Syndrome'.

The `SRSTR` table in the Semantic Network describes the structure of the network. This table lists a range of different relations between semantic types, including hierarchical relations (`is_a`) and non hierarchical relations (e.g. `result of`, `associated with` and `co-occurs with`). For example, the semantic types 'Disease or Syndrome' and 'Pathologic Function' are connected via the `is_a` relation in this table.

The **MetaMap** program (Aronson, 2001) maps biomedical text to concepts in the Metathesaurus. The semantic type for each concept mapping is also returned. MetaMap employs a knowledge intensive approach that uses the Specialist Lexicon in combination with lexical and syntactic analysis to identify noun phrases in text. Matches between noun phrases and Metathesaurus concepts are computed by generating lexical variations and allowing partial matches between the phrase and concept. The possible UMLS concepts are assigned scores based on the closeness of the match between the input noun phrase and the target concept. Fig. 1 shows this mapping for the phrase "*tissues are often cold*". This example shows that MetaMap returns a single CUI for two words (*tissues* and *often*) but also returns multiple CUIs with equal scores for *cold* (C0234192, C0009443 and C0009264). Weeber et al. (2001) estimated that around 11% of the phrases in Medline abstracts are mapped onto multiple CUIs.