Concept-graph based Biomedical Automatic Summarization using Ontologies

Laura Plaza Morales
Alberto Díaz Esteban
Pablo Gervás
Universidad Complutense de Madrid
C/Profesor José García Santesmases, s/n, Madrid 28040, Spain
lplazam@pas.ucm.es, albertodiaz@fdi.ucm.es,pgervas@sip.ucm.es

Abstract

One of the main problems in research on automatic summarization is the inaccurate semantic interpretation of the source. Using specific domain knowledge can considerably alleviate the problem. In this paper, we introduce an ontology-based extractive method for summarization. It is based on mapping the text to concepts and representing the document and its sentences as graphs. We have applied our approach to summarize biomedical literature, taking advantages of free resources as UMLS. Preliminary empirical results are presented and pending problems are identified.

1 Introduction

In recent years, the amount of electronic biomedical literature has increased explosively. Physicians and researchers constantly have to consult up-to date information according to their needs, but the process is time-consuming. In order to tackle this overload of information, text summarization can undoubtedly play a role.

Simultaneously, a big deal of resources, such as biomedical terminologies and ontologies, have emerged. They can significantly benefit the development of NLP systems, and in particular, when used in automatic summarization, they can increase the quality of summaries.

In this paper, we present an ontology-based extractive method for the summarization of biomedical literature, based on mapping the text to concepts in UMLS and representing the document and its sentences as graphs. To assess the importance of the sentences, we compute the centrality of their concepts in the document graph.

2 Previous Work

Traditionally, automatic summarization methods have been classified in those which generate *extracts* and those which generate *abstracts*. Although human summaries are typically abstracts, most of existing systems produce extracts.

Extractive methods build summaries on a superficial analysis of the source. Early summarization systems are based on simple heuristic features, as the position of sentences in the document (Brandow et al., 1995), the frequency of the words they contain (Luhn, 1958; Edmundson, 1969), or the presence of certain cue words or indicative phrases (Edmundson, 1969). Some advanced approaches also employ machine learning techniques to determine the best set of attributes for extraction (Kupiec et al., 1995). Recently, several graph-based methods have been proposed to rank sentences for extraction. LexRank (Erkan and Radev, 2004) is an example of a centroid-based method to multi-document summarization that assess sentence importance based on the concept of eigenvector centrality. It represents the sentences in each document by its *tf*idf* vectors and computes sentence connectivity using the cosine similarity. Even if results are promising, most of these approaches exhibit important deficiencies which are consequences of not capturing the semantic relations between terms (synonymy, hyperonymy, homonymy, and *co-occurs* and *associated-with* relations).

We present an extractive method for summarization which attempts to solve this deficiencies. Unlike researches conducted by (Yoo et al., 2007; Erkan and Radev, 2004), which cluster sentences to identify shared topics in multiple documents, in this work we apply clustering to identify groups of concepts closely related. We hypothesize that

each cluster represents a *theme* or topic in the document, and we evaluate three different heuristics to ranking sentences.

3 Biomedical Ontologies. UMLS

Biomedical ontologies organize domain concepts and knowledge in a system of hierarchical and associative relations. One of the most widespread in NLP applications is UMLS[1] (Unified Medical Language System). UMLS consists of three components: the *Metathesaurus*, a collection of concepts and terms from various vocabularies and their relationships; the *Semantic Network*, a set of categories and relations used to classify and relate the entries in the Metathesaurus; and the *Specialist Lexicon*, a database of lexicographic information for use in NLP. In this work, we have selected UMLS for several reasons. First, it provides a mapping structure between different terminologies, including MeSH or SNOMED, and thus allows to translate between them. Secondly, it contains vocabularies in various languages, which allows to process multilingual information.

4 Summarization Method

The method proposed consists of three steps. Each step is discussed in detail below. A preliminary system has been implemented and tested on several documents from the corpus developed by *BioMed Central*[2].

As the preprocessing, text is split into sentences using GATE[3], and generic words and high frequency terms are removed, as they are not useful in discriminating between relevant and irrelevant sentences.

4.1 Graph-based Document Representation

This step consists in representing each document as a graph, where the vertices are the concepts in UMLS associated to the terms, and the edges indicate the relations between them. Firstly, each sentence is mapped to the UMLS Metathesaurus using *MetaMap* (Aronson, 2001). MetaMap allows to map terms to UMLS concepts, using n-grams for indexing in the ULMS Metathesaurus, and performing disambiguation to identify the correct concept for a term. Secondly, the UMLS concepts are extended with their hyperonyms. Figure 1 shows the graph for sentence *"The goal of the trial was to assess cardiovascular mortality and morbidity for stroke, coronary heart disease and congestive heart failure, as an evidence-based guide for clinicians who treat hypertension."*

Next, each edge is assigned a weight, which is directly proportional to the deep in the hierarchy at which the concepts lies (Figure 1). That is to say, the more specific the concepts connected are, the more weight is assigned to them. Expression (1) shows how these values are computed.

$$\frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} = \frac{|\beta|}{|\alpha|} \quad (1)$$

where α is the set of all the parents of a concept, including the concept, and β is the set of all the parents of its immediate higher-level concept, including the concept.

Finally, the sentence graphs are merged into a document graph, enriched with the *associated-with* relations between the semantic types in UMLS corresponding to the concepts (Figure 1). Weights for the new edges are computed using expression (1).

4.2 Concept Clustering and Theme Recognition

The second step consists of clustering concepts in the document graph, using a *degree-based method* (Erkan and Radev, 2004). Each cluster is composed by a set of concepts that are closely related in meaning, and can be seen as a theme in the document. The most central concepts in the cluster give the sufficient and necessary information related to its theme. We hypothesize that the document graph is an instance of a *scale-free network* (Barabasi, 1999). Following (Yoo et al., 2007), we introduce the *salience* of vertices. Mathematically, the salience of a vertex (v_i) is calculated as follows.

$$salience(v_i) = \sum_{e_j | \exists v_k \wedge e_j conecta(v_i, v_k)} weight(e_j) \quad (2)$$

Within the set of vertices, we select the *n* that present the higher salience and iteratively group them in *Hub Vertex Sets* (HVS). A HVS

[1] NLM Unified Medical Language System (UMLS). URL: http://www.nlm.nih.gov/research/umls

[2] BioMed Central: http://www.biomedcentral.com/

[3] GATE (Generic Architecture for Text Engineering): http://gate.ac.uk/

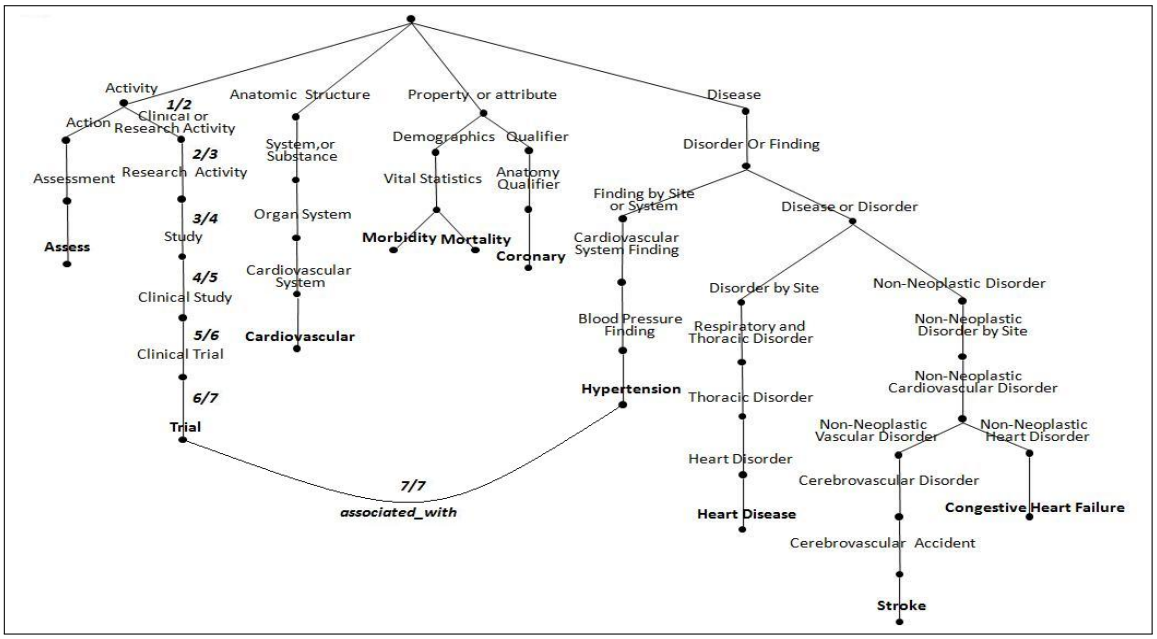

Figure 1: Sentence graph

represents a group of vertices that are strongly related to each other. The remaining vertices are assigned to that cluster to which they are more connected.

Finally, we assign each sentence to a cluster. To measure the similarity between a cluster and a sentence graph, we use a vote mechanism (Yoo et al., 2007). Each vertex (v_k) of a sentence (O_j) gives to each cluster (C_i) a different number of votes ($p_{i,j}$) depending on whether the vertex belongs to HVS or non-HVS (3).

$$similarity(C_i, O_j) = \sum_{v_k | v_k \in O_j} w_{k,j} \qquad (3)$$

where $\begin{cases} w_{k,j}=0 \; si \; v_k \notin C_i \\ w_{k,j}=1.0, si \; v_k \in HVS(C_i) \\ w_{k,j}=0.5, si \; v_k \notin HVS(C_i) \end{cases}$

4.3 Sentence Selection

The last step consists of selecting significant sentences for the summary, based on the similarity between sentences and clusters. We investigated three alternatives for this step.

- **Heuristic 1:** For each cluster, the top n_i sentences are selected, where n_i is proportional to its size.
- **Heuristic 2:** We accept the hypothesis that the cluster with more concepts represents the main theme in the document, and select the top N sentences from this cluster.
- **Heuristic 3:** We compute a single score for each sentence, as the sum of the votes assigned to each cluster adjusted to their sizes, and select the N sentences with higher scores.

5 Results and Evaluation

In order to evaluate the method, we analyze the summaries generated by the three heuristics over a document[4] from the BioMed Central Corpus, using a compression rate of 20%. Table 1 shows the sentences selected along with their scores.

Although results are not statistically significant, they show some aspects in which our method behaves satisfactorily. Heuristics 1 and 3 extract sentence 0, and assign to it the higher score. This supports the positional criterion of selecting the first sentence in the document, as the one that contains the most significant information. Sentence 58 represents an example of sentence, situated at the end, which gathers the conclusions of the author. In general, these sentences are highly informative. Sentence 19, in turn, evidences how the method systematically gives preference to long sentences. Moreover, while summaries by heuristics 1 and 3

[4] BioMed Central: www.biomedcentral.com/content/download/xml/cvm-2-6-254.xml

Sentences	0	4	19	58	7	28	25	20	21	8	43	15
Heuristic 1	99.0	20.0	19.0	18.5	17.0	16.5	16.0	15.5	15.5	13.5	13.5	12.0
Heuristic 2	19.0	16.5	15.5	12.5	12.0	10.5	9.0	9.0	7.5	7.0	7.0	7.0
Heuristic 3	98.8	18.7	17.9	16.3	15.3	14.5	13.4	13.0	13.0	12.7	12.7	12.2

Table 1: Results

have a lot of sentences in common (9 out of 12), heuristic 2 generates a summary considerably different and ignores important topics in the document. Finally, we have compared these summaries with the author's abstract. It can be observed that heuristics 1 and 3 cover all topics in the author's abstract (see sentences 0, 4, 15, 17, 19, 20 and 25). As far as heuristic 2 is concerned, it does not cover adequately the information in the abstract.

6 Conclusions and Future Work

In this paper we introduce a method for summarizing biomedical literature. We represent the document as an ontology-enriched scale-free graph, using UMLS concepts and relations. This way we get a richer representation than the one provided by a vector space model. In section 5 we have evaluated several heuristics for sentence extraction. We have determined that heuristic 2 does not cover all relevant topics and selects sentences with a low relative significance. Conversely, heuristics 1 and 3, present very similar results and cover all important topics.

Nonetheless, we have identified several problems and some possible improvements. Firstly, as our method extracts whole sentences, long ones have higher probability to be selected, because they contain more concepts. The alternative could be to normalise the sentences scores by the number of concepts. Secondly, concepts associated with general semantic types in UMLS, as *functional concept*, *temporal concept*, *entity* and *language*, could be ignored, since they do not contribute to distinguish what sentences are significant.

Finally, in order to formally evaluate the method and the different heuristics, a large-scale evaluation on the BioMed Corpus is under way, based on computing the ROUGE measures (Lin, 2004).

Acknowledgements

This research is funded by the Ministerio de Educación y Ciencia (TIN2006-14433-C02-01), Universidad Complutense de Madrid and Dirección General de Universidades e Investigación de la Comunidad de Madrid (CCG07-UCM/TIC 2803).

References

Aronson A. R. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. 2001. *In Proceedings of American Medical Informatics Association*.

Barabasi A.L. and Albert R. Emergence of scaling in random networks. 1999. *Science*,286–509.

Brandow R. and Mitze K. and Rau L. F. Automatic Condensation of Electronic Publications by Sentence Selection. 1995. *Information Processing and Management*,5(31):675–685.

Edmundson H.P. New Methods in Automatic Extracting. 1969. *Journal of the Association for Computing Machinery*,2(16):264–285.

Erkan G. and Radev D. R. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. 2004. *Journal of Artificial Intelligence Research (JAIR)*,22:457–479.

Kupiec J. and Pedersen J.O. and Chen F. A Trainable Document Summarizer. 1995. *In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*,68–73.

Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. 2004. *In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*.

Luhn H.P. The Automatic Creation of Literature Abstracts. 1958. *IBM Journal of Research Development*,2(2):159–165.

Sparck-Jones K. Automatic Summarizing: Factors and Directions. 1999. *I. Mani y M.T. Maybury, Advances in Automatic Text Summarization*. The MIT Press.

Yoo I. and Hu X. and Song I.Y. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. 2007. *BMC Bioinformatics*,8(9).