Generating Links to Background Knowledge for Medical Content*

Jiyin He
Centrum Wiskunde en Informatica
Science Park 123, 1098XG
Amsterdam, the Netherlands
J.He@cwi.nl

Maarten de Rijke
University of Amsterdam
Science Park 904, 1098XH
Amsterdam, the Netherlands
derijke@uva.nl

Merlijn Sevenster
Philips Research
High Tech Campus 34,5656AA
Eindhoven, the Netherlands
merlijn.sevenster@philips.com

ABSTRACT

Automatically annotating texts with background information has recently received much attention. In this note, we outline a case study in automatically generating links from narrative radiology reports to Wikipedia. Such links are meant to help users understand the medical terminology and thereby increase the value of the reports. We discuss our findings throughout this study and the open questions left to be explored. In particular, we discuss potential extensions to the general medical domain.

1. INTRODUCTION

Hypertext links help users navigate to pertinent information, also when they are not aware that the information exists or when they they are incapable of articulating an appropriate search query. Many types of links exist, e.g., categorical links (such as the navigation of a website), links to related items (such as linking events in news along a timeline), links to "similar items" (in book reviews, or in online shopping environments), etc. We shall focus on explanatory links — that is, the target of a link provides definitions or background information for the source of the link —, and we shall do so in the medical domain. Such links are motivated by the typical scenario in which a patient encounters medical terms in his/her medical report that are beyond his/her comprehension, which impedes him/her from understanding the report. To oversee his/her health status and treatment options, the patient searches the web for formation, see, e.g., "What is lacunar infarct? It showed up on a CT but I cannot find anything about it on the medical sites?"[1] It is envisioned that by automatically identifying medical terms and explaining them through a link to a knowledge resource that is accessible and understandable by non-experts, which may result in improved expert-patient communication and increased patient empowerment.

Automatically generating links to Wikipedia has received much attention in recent years. Most of the studies are interested in solving a generic problem (e.g., developing an automatic link generation approach using Wikipedia as training material [2, 3]) and aim to generate links for texts on general topics, e.g., news articles.

In this note we sketch why the previous approaches do not work in our domain of interest; we review the case study [1] in which a new link generation system was proposed; and we touch on directions for future research.

2. AUTOMATIC LINK GENERATION FOR RADIOLOGY REPORTS

A radiology report gives a narrative account of the radiologist's findings, diagnoses and recommendations for followup actions. It is the principal means of communication between radiologists and referring clinicians such as surgeons and general practitioners.

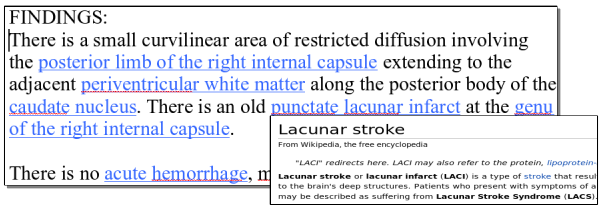

Figure 1: Medical phrases linked to Wikipedia.

In [1], a case study was conducted to link anatomy and diagnosis terms found in radiology report to Wikipedia pages. Figure 1 shows an excerpt from a radiology report, with medical terms requiring explanation highlighted, together with a snippet from a Wikipedia page describing one of the highlighted medical terms.

Two two state-of-the-art systems, Wikify! [2] and Wikipedia Miner [3], were evaluated on a test collection of 860 radiology reports that were annotated and aggregated manually by domain experts with links to Wikipedia concepts. Both systems are assumable domain independent, and were trained using existing Wikipedia links.

It was found that, first, neither system yields satisfactory results on our test collection. This was due to two properties of medical phrases. (1) Medical phrases are often syntactically regular: they are mostly noun phrases with one or more modifiers (e.g., adjectives); and (2) while syntactically regular, medical phrases often have a complicated semantic structure, due to the presence of multiple modifiers as well as conjunctions of concepts within a single phrase. The latter property is the major reason why both systems under perform, as Wikipedia concepts are often relatively simple, e.g., consist of a single concept or concepts without modifiers.

An automatic link generation approach was proposed that addresses these properties. A sequential labeling approach was used

*This note is based on the study reported in [1].
[1]http://au.answers.yahoo.com/question/index?qid=1006032312565

MEDEX'11 October 28, 2011, Glasgow
.

with syntactic features to identify anchor texts in order to exploit syntactic regularities in medical terminology. This was combined with a sub-anchor approach to target finding, which is aimed at coping with the complex semantic structure of medical phrases. The proposed system was shown to effectively improve over the two state-of-the-art systems.

Second, it was found that automatic link generation systems tend to achieve better performance in recognizing and finding targets for frequent anchor texts than for infrequent anchor texts. In order to achieve robust performance, it is therefore important that a system is effective when dealing with infrequent anchor texts.

3. DISCUSSION AND OPEN QUESTIONS

The reviewed case study has uncovered some of the properties of the problem of linking medical terminology to background knowledge; the proposed approach has shown to be effective. It opened up a number of directions for future research.

3.1 Beyond Wikipedia

In our study, we have focused on linking to Wikipedia and following the state-of-the-art automatic (Wikipedia) link generation studies, we used a pure data-driven approach. Other medical resources exist besides Wikipedia, which may provide valuable training data for link generation. For example, at the National Library of Medicine (NLM), domain experts index (or annotate) biomedical articles from MedLine with terms from Medical Subject Headings (MeSH). This annotation can be utilized as training data for linking medical concepts in free text to the MeSH terms. Moreover, unlike Wikipedia in which concepts are organized in a network structure and a link indicates an explanatory relation between two concepts, MeSH organizes its terms in a hierarchical structure. The semantics of the relations that constitute this hierarchy are clearly defined, and can thus be leveraged for hierarchical reasoning. This opens up the opportunity to enhance the data-driven approach of [1] with symbolic approaches that make inferences on the basis of the ontology structure.

Continuing this line of thinking, we note that many medical Wikipedia pages are labeled with concepts from medical ontologies such as MeSH and the International Classification of Diseases (ICD). We can also use these concepts to make inferences in a background ontology to retrieve potentially more relevant Wikipedia pages.

3.2 Linking terms in general medical content

Our case study was carried out using neuroradiology reports, which share many similarities with other types of medical reports (e.g., reports of other radiology specialties, pathology reports or discharge letters). It would be interesting to apply and extend the link generation methods to such types of reports as well.

We can also envision generating links from medical data other than medical reports, such as user generated content on the Internet. For example, there are online forums (e.g., Yahoo! Answers) where people can discuss their health situations, receiving responses from experts and laymen alike. Automatically linking terminologies in such discussions to background knowledge can be helpful, as we have seen that occasionally some participants simply contribute by copying and pasting definitions of terminologies found in a medical resource. This indicates that automatically generating such links can save the effort of looking for definitions manually. In addition, it may help to steer the user to medical data [4] that is known to be reliable.

As shown in the case study, the state-of-the-art approaches do not work effectively on linking radiology reports due to the structure mismatch between Wikipedia links and annotated links in the radiology reports. While extending the approach to the aforementioned medical content, adaptation may be required to maintain accuracy.

For example, the risk of generalizing from a specific domain to a more general domain is that concepts in a general domain are more likely to be ambiguous compared to concepts in the specific domain. The word "ventricle," for instance, is ambiguous between a space in the heart and an area in the brain. Therefore, if we widen our scope from neuroradiology reports to cardiac reports, such ambiguities should be accounted for.

On the other hand, when extending our work from medical reports to user generated medical content, the opposite of "disambiguation" is needed. For instance, on forums people tend to use laymen's names for diseases instead of using their scientific names, while in medical reports scientific names and standardized codes are used (e.g., ICD codes). So efforts need to be made to normalize medical terminology.

3.3 User perspective

In our previous study, we focused on the precision and recall of re-generating the links created by domain experts and evaluated the system performance in an abstracted setting. However, what types of links are most appreciated by patients? For instance, links that explain medical conditions, physiological causes, or treatment options? Moreover, among non-experts, individuals possess different levels of medical knowledge, and may therefore require personalized content. Besides patients' interests, it may be equally valuable to address the interests of medical professionals. For them it may be valuable to link phrases to expert encyclopedias and scientific literature such as PubMed. It is of eminent importance to track how people's medical information consumption changes when they are exposed to automatically generated links, compare [4].

4. CONCLUSION

We discussed a recent case study on generating links from radiology reports to a general knowledge source, Wikipedia. Both lessons learned and future research directions were identified.

Acknowledgments

This research was supported by the EU's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the Fish4Knowledge project and the PROMISE Network of Excellence, funded and co-funded by the 7th Framework Programme of the European Commission, grant agreement nr 257024 and nr 258191, respectively, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

5. REFERENCES

[1] J. He, M. de Rijke, M. Sevenster, R. van Ommering, and Y. Qian. Generating links to background knowledge: a case study in annotating radiology reports. In *CIKM'11*, 2011.

[2] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07*, pages 233–242, 2007.

[3] D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08*, pages 509–518, New York, NY, USA, 2008.

[4] R. W. White and E. Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Trans. Inf. Syst.*, 27:23:1–23:37, November 2009. ISSN 1046-8188.