Exploring the Use of NLP in the Disclosure of Electronic Patient Records

David Hardcastle
Faculty of Mathematics and Computing
The Open University
d.w.hardcastle@open.ac.uk

Catalina Hallett
Faculty of Mathematics and Computing
The Open University
c.hallett@open.ac.uk

Abstract

This paper describes a preliminary analysis of issues involved in the production of reports aimed at patients from Electronic Patient Records. We present a system prototype and discuss the problems encountered.

1 Introduction

Allowing patient access to Electronic Patient Records (EPR) in a comprehensive format is a legal requirement in most European countries. Apart from this legal aspect, research shows that the provision of clear information to patients is instrumental in improving the quality of care (Detmer and Singleton, 2004). Current work on generating explanations of EPRs to patients suffer from two major drawbacks. Firstly, existing report generation systems have taken an intuitive approach to the generation of explanation: there is no principled way of selecting the information that requires further explanation. Secondly, most work on medical report generation systems has concentrated on explaining the structured part of an EPR; there has been very little work on providing automatic explanations of the narratives (such as letters between health practitioners) which represent a considerable part of an EPR. Attempting to rewrite narratives in a patient-friendly way is in many ways more difficult than providing suggestions for natural language generation systems that take as input data records. In narratives, ambiguity can arise from a combination of aspects over which NLG systems have full control, such as syntax, discourse structure, sentence length, formatting and readability.

This paper introduces a pilot project that attempts to address this gap by addressing the following research questions:

1. Given the text-based part of a patient record, which segments require explanation before being released to patients?
2. Which types of explanation are appropriate for various types of segment?
3. Which subparts of a segment require explanation?

The prototype system correctly selects the segments that require explanation, but we have yet to solve the problem of accurately identifiying the features that contribute to the "expertness" of a document. We discuss the underlying issues in more detail in section 3 below.

2 Feature identification method

To identify a set of features that differentiate medical expert and lay language, we compared a corpus of expert text with a corpus of lay texts. We then used the selected features on a corpus of narratives extracted from a repository of Electronic Patient Records to attempt to answer the three questions posed above. First, paragraphs that contain features characteristic to expert documents are highlighted using a corpus of patient information leaflets as a background reference. Second, we prioritise the explanations required by decomposing the classification data. Finally, we identify within those sections the features that contribute to the classification of the section as belonging to the expert register, and provide suggestions for text simplification.

2.1 Features

The feature identification was performed on two corpora of about 200000 words each: (a) an expert corpus, containing clinical case studies and medical manuals produced for doctors and (b) a lay corpus, containing patient testimonials and informational materials for patients. Both corpora were

BioNLP 2007: Biological, translational, and clinical language processing, pages 161–162,
Prague, June 2007. ©2007 Association for Computational Linguistics

sourced from a variety of online sources. In comparing the corpora we considered a variety of features in the following categories: medical content, syntactic structure, discourse structure, readability and layout. The features that proved to be best discriminators were the frequency of medical terms, readability indices, average NP length and the relative frequency of loan words against English equivalents[1]. The medical content analysis is based on the MeSH terminology (Canese, 2003) and consists of assessing: (a) the frequency of MeSH primary concepts and alternative descriptions, (b) the frequency of medical terms types and occurences and (c) the frequency of MeSH terms in various top-level categories. The readability features consist of two standard readability indices (FOG and Flesch-Kincaid). Although some discourse and layout features also proved to have a high discriminatory power, they are strongly dependent on the distribution medium of the analysed materials, hence not suitable for our analysis of EPR narratives.

2.2 Analysing EPR narratives

We performed our analysis on a corpus of 11000 narratives extracted from a large repository of Electronic Patient Records, totalling almost 2 million words. Each segment of each narrative was then assessed on the basis of the features described above, such as Fog, sentence length, MeSH primary concepts etc. We then smoothed all of the scores for all segments for each feature forcing the minimum to 0.0, the maximum to 1.0 and the reference corpus score for that feature to 0.5. This made it possible to compare scores with different gradients and scales against a common baseline in a consistent way.

3 Evaluation and discussion

We evaluated our segment identification method on a set of 10 narratives containing 27 paragraphs, extracted from the same repository of EPRs . The segment identification method proved succesful, with 26/27 (96.3%) segments marked correctly are requiring/not requiring explanation. However, this only addresses the first of the three questions set out above, leaving the following research questions open to further analysis.

Quantitative vs qualitative analysis
Many of the measures that discriminate expert from lay texts are based on indicative features; for example complex words are indicative of text that is difficult to read. However, there is no guarantee that individual words or phrases that are indicative are also representative - in other words a given complex word or long sentence will contribute to the readability score of the segment, but may not itself be problematic. Similarly, frequency based measures, such as a count of medical terminology, discriminate at a segment level but do not entail that each occurrence requires attention.

Terminology
We used the MeSH terminology to analyse medical terms in patient records, however (as with practically all medical terminologies) it contains many non-expert medical terms. We are currently investigating the possibility of mining a list of expert terms from MeSH or of making use of medical-lay aligned ontologies.

Classification
Narratives in the EPR are written in a completely different style from both our training expert corpus and the reference patient information leaflets corpus. It is therefore very difficult to use the reference corpus as a threshold for feature values which can produce good results on the corpus of narratives, suggesting that a statistical thresholding technique might be more effective.

Feature dependencies
Most document features are not independent. Therefore, the rewriting suggestions the system provides may themselves have an unwanted impact on the rewritten text, leading to a circular process for the end-user.

References

Kathi Canese. 2003. New Entrez Database: MeSH. *NLM Technical Bulletin*, March-April.

D. Detmer and P. Singleton. 2004. The informed patient. Technical Report TIP-2, Judge Institute of Management, University of Cambridge, Cambridge.

Noemi Elhadad. 2006. Comprehending technical texts: Predicting and defining unfamiliar terms. In *Proceeding of AMIA'06*, pages 239–243.

[1] An in-depth analysis of unfamiliar terms in medical documents can be found in (Elhadad, 2006)