Iterative refinement of SemLink to enhance patient readability of discharge summaries

Mehnaz Adnan a, Jim Warren b and Martin Orr c

a Department of Computer Engineering, Sir Syed University of Engineering and Technology, Pakistan  
b Department of Computer Science, University of Auckland  
c School of Population Health, University of Auckland  
c Waitemata District Health Board, Auckland, New Zealand

Abstract. A Discharge Summary contains vocabulary that is difficult to understand for health consumers. We used iterative refinements in developing a system, SemLink, which dynamically generate synonyms and hyperlinks to appropriate Internet resources for difficult terms in discharge summary text to make the text more comprehensible to consumers. This paper describes our iterative refinement protocol to enhance the semantic annotation and dynamic hyperlinking algorithms to link topic-specific web pages for difficult terms found occurring in Discharge Summary text.

Keywords. Discharge Summary, Semantic Annotation, Hyperlink, Readability

1. Introduction

Health consumers use information in their Discharge Summaries (DS) for later consultation and self-care management [1]. However, these summaries contain vocabulary that are difficult to understand for an average consumer[2]. Efforts to improve consumer-friendliness of clinical text have focused on providing links to educational materials in Pap smear reports [3] and text translation [4]. We are using automatic semantic annotationsto develop a prototype system, SemLink, which provides synonyms and dynamic hyperlinks for difficult text in Electronic Discharge Summaries (EDSs) to aid consumers in reading their DS online[5]. SemLink was implemented in iterative refinements with interim evaluations. In our previous phase we provided semantic annotations for difficult terms based on CHV (the Open Access Collaborative’s Consumer Health Vocabulary, http://consumerhealthvocab.org/) and for commonly used abbreviations. The hyperlinks are provided dynamically by automatically selecting the best link from among the first search results page offered by MedlinePlus, WebMD, Healthline, Patient UK, Medsafe and Yahoo! Health. This system was evaluated by automatically measuring the efficiency of the dynamic hyperlink generation in terms of precision, recall and readability scores of the hyperlinked Web pages. The result was promising, but in need of refinement[6]. In this study we describe our iterative refinement protocol, and resulting enhanced semantic annotation approach for identification of difficult terms in EDS text and final dynamic hyperlinking algorithm.
2. Iterative Refinement of SemLink

We retrieved 50 EDSs each from Emergency, Medicine, Surgery and Assessment, Treatment & Rehabilitation (AT&R) Services collected from a total of 62,674 EDS generated in hospitals of Waitemata District Health Board (WDHB) during the period of June 2007 to July 2008. We performed eight iterations: two on clinical management sections provided by the National E-Health Transition Authority Australia\(^1\) and six on clinical management sections from randomly selected EDSs from our sample data for analysis in an iterative refinement protocol. Two experts, one experienced in health informatics and one consultant physician with health informatics background [2\(^{nd}\) and 3\(^{rd}\)authors] analysed the annotated clinical management sections and provided their feedback on the annotations and the hyperlinks provided by SemLink. The total numbers of words in these sections were 1434 in which 214 annotations and hyperlinks were analyzed for their quality in terms of correctness and topic relevancy. After each iteration the identified issues were fixed in the system. The key findings which were incorporated during iterative refinement of SemLink are as follows:

**Title Term Frequency (TF):** During the first iteration, it was observed that a page title TF strictly greater than 0.5 was an effective cut-off for annotation of the term. For example, for the term ‘leukemia’, MedlinePlus retrieves two Web pages: one with a title ‘leukemia’ (TF=1 apropos the term being annotated), and the other with the title ‘childhood leukemia’ (TF=0.5). As with this example, pages with TF exactly 0.5 are often significantly divergent from the annotated concept.

**Dictionary Interfaces:** During the second iteration, we found that many CHV difficult terms were not provided with a hyperlink. After this iteration, the list of search portals was extended and online dictionary interfaces were also included to provide definitions for those terms for which no title-specific page could be selected from any of the other portals used. Lists of included search portals and dictionary interfaces are shown in Table 1 and Table 2, respectively.

**UMLS Mapping:** In the first and second iterations we found using CHV alone is not sufficient to annotate concepts in EDS text. Therefore we decided to first use UMLS for concept mapping in EDS text and to concentrate only for those UMLS concepts having CHV difficult term or abbreviation encapsulated in the mapped UMLS concept.

**Non-Difficult Terms:** We built a list of words for which no readability support is needed. This list includes stop words available from [7], the words in our corpus which we feel are commonly understood by a native English speakers (e.g. ‘forearm’ is scored as difficult by CHV, possibly due to low frequency in the training set used for its creation, but we felt best left un-annotated) and acronyms having frequency less than twenty in our EDS corpus. This list serves to filter the SemLink annotations.

**Expansion of WDHB Abbreviations:** During second and third iteration, we came across acronyms which are commonly used in our corpus but not included in the abbreviation list provided by WDHB and incorrect mapping of two letter acronyms by UMLS. To address this issue, we extended abbreviation list with those abbreviations in our EDS corpus which are commonly used in WDHB communication. Two letter acronyms which are present twenty or more times in our EDS corpus are also included in the abbreviation list.

\(^1\)http://www.nehta.gov.au
UMLS Semantic Types: In the fourth iteration, many terms are found which are difficult but not annotated, for example ‘tachypnoea’. In this iteration we decided to use the UMLS semantic network and include the relevant UMLS semantic types for readability support in SemLink. The included semantic types are shown in Figure 1. In addition, semantic types qualifier values, contextual qualifier, qualitative concept and temporal concept were omitted due to their lack of task relevancy and ambiguity.

3. Enhanced System Design

3.1. Semantic Annotations in SemLink

The semantic annotations are created based on the Terminology ontology as shown in Figure 1. Three main classes, CHV_Difficult_Term, UMLS_Concept and Abbreviation are created to model CHV difficult terms, difficult UMLS concepts and WDHB abbreviations, respectively. The ontology relations hasCHVDifficultTerm and hasAbbreviation are used in such a way that a UMLS concept is considered difficult if it contains any instance of CHV difficult terms or abbreviation in it.

![Figure 1. Terminology ontology in SemLink](image)

The HITEx UMLS concept finder, CHV, and the WDHB provided abbreviation list, were used to annotate the instances of UMLS_Concept, CHV_Difficult_Term and Abbreviation in EDS text using GATE(https://www.gate.ac.uk/) for natural language processing (NLP). These annotations are embedded in a PDF and an example annotated phrase underlined in PDF document is shown in Figure 2. In this text, when the user clicks on the underlined term, the interface shows its preferred name in tooltip and opens the related hyperlink in a new browser window as selected by the algorithm described in the next section. A sample clinical management section with embedded annotations in PDF is shown in the first author’s thesis[5].

![Figure 2. Semantic annotations embedded in PDF format](image)

3.2. Final Hyperlinking Algorithm

The final hyperlink generation algorithm as shown in Figure 3 leverages 11 search engines and three online dictionaries as listed in Table 1 and Table 2 respectively. This algorithm is composed of two main components; a portal page retrieval component and a definition retrieval component. The portal page retrieval component helps in selecting a topically relevant and readable Web resource provided by the search engines listed in Table 1 by offering an efficient search and result set filtering mechanism. The definition retrieval module provides a definition of the annotated term or phrase in cases where no title-specific resource is found by the portal page retrieval module.
A PHP script was written to implement this algorithm. Initially the algorithm takes GATE annotations for the annotated term or phrase to be hyperlinked as input. These include preferred term, ontology class, UMLS Name, and UMLS Concept Unique Identifier (CUI). Synonyms were extracted for each annotated concept using its CUI from UMLS and CHV databases.

Table 1. Consumer health search engines used in SemLink

<table>
<thead>
<tr>
<th>Search Engine Name</th>
<th>Available From</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedlinePlus</td>
<td><a href="http://medlineplus.gov/">http://medlineplus.gov/</a></td>
</tr>
<tr>
<td>Healthline</td>
<td><a href="http://www.healthline.com/">http://www.healthline.com/</a></td>
</tr>
<tr>
<td>WebMD</td>
<td><a href="http://www.webmd.com/">http://www.webmd.com/</a></td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td><a href="http://mayoclinic.com/health/">http://mayoclinic.com/health/</a></td>
</tr>
<tr>
<td>Patient UK</td>
<td><a href="http://www.patient.co.uk/">http://www.patient.co.uk/</a></td>
</tr>
<tr>
<td>Medsafe</td>
<td><a href="http://www.medsafe.govt.nz/Consumers">http://www.medsafe.govt.nz/Consumers</a></td>
</tr>
<tr>
<td>Family Doctor</td>
<td><a href="http://www.familydoctor.co.nz/">http://www.familydoctor.co.nz/</a></td>
</tr>
<tr>
<td>Health.com</td>
<td><a href="http://www.health.com/">http://www.health.com/</a></td>
</tr>
<tr>
<td>MedTerms</td>
<td><a href="http://www.medterms.com/">http://www.medterms.com/</a></td>
</tr>
</tbody>
</table>

Table 2. Online dictionary interfaces used in SemLink

<table>
<thead>
<tr>
<th>Dictionary Name</th>
<th>Available From</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedlinePlus</td>
<td><a href="http://www.merriamwebster.com/medlineplus/">http://www.merriamwebster.com/medlineplus/</a></td>
</tr>
</tbody>
</table>

The portal page retrieval component first executes the search for the preferred term in MedlinePlus. The title matching module loads the search results found by MedlinePlus and extracts the title of pages to calculate Term Frequency (TF) for each page title. The TF is used as a metric to gauge which Web pages are topically relevant to the annotated term or its synonyms, in terms of matching words. Stop words are removed from both the Web page title and synonyms using a widely-available stop word list[7], and both are stemmed using Porter stemming algorithm. For each synonym, the frequency for each stemmed word in the page title was found, and if the title length is greater than or equal to the synonym length, its TF score was calculated as:

\[ \sum_{k=1}^{n} \frac{f_{kn}}{N} \]

In this equation, TF_{mn} is the term frequency of the m_{th} synonym, F_k is the frequency of the k_{th} stemmed word of m_{th} synonym in the page title and N is equal to the total number of stemmed words in the page title. All the URLs with TF score greater than 0.5 are indexed as high term frequency results and are taken to be closely related URLs for the annotated term. In this way the title matcher module increases the precision of the hyperlinked pages, by preventing unrelated Web pages from being associated with the annotated segment in the text. For the cases, where multiple URLs are returned with TF greater than 0.5, a readability optimization module is added to select the web page having better readability index. This module loads the pages having TF greater than 0.5 and computes the SMOG score [8] of each Web page and select the page having TF greater than 0.5 and minimum SMOG score.

In cases where the title matching module does not return a MedlinePlus index URL then the same procedure iterates through the rest of the health portals listed in Table 1. Google Web search API was used to search in these portals by restricting the search to the listed portal. For each included portal, the first four URLs returned by the Google
was retrieved and title specific web selected in the same fashion as selected for MedlinePlus retrieved URLs. If no title specific page found, the definition retrieval component uses the dictionary interfaces in the same order as listed in Table 2 to retrieve definition of the annotated term. The preferred term was used to retrieve the definition of the preferred term of the annotated term via the Yahoo! Kids and MedlinePlus dictionary interface. If no definition found from these dictionary interfaces, a WordNet definition for the preferred term of the annotated term is retrieved.

Figure 3. Flowchart of the final hyperlink generation algorithm.

4. Discussion

The need to improve the readability of DS text for consumers is growing as patients access these summaries for self-care management. We adopted an iterative refinement protocol to enhance the design of SemLink which embeds semantic annotations in the form of synonyms and dynamic hyperlinks to selected internet resources to provide readability assistance for consumers in EDSs. The iterative refinements led to changes in the previous design of SemLink[6], including use of UMLS for term identification and UMLS semantic types to focus hyperlinking, stricter criteria for annotation, refinement of our heuristics for acronyms and expansion of the set of internet resources used.

Automatic semantic annotations and dynamic hyperlinking to uncontrolled web resources is inherently challenging. The experts’ feedback reported many problems related to automatic semantic annotations. For example, UMLS mapped ‘ACH’ to acetylcholinesteratne instead of ‘Auckland City Hospital’ in our EDS corpus. In addition many difficult terms are not provided with hyperlinks as per our previous algorithm. These issues were fixed by introducing UMLS and by extending the list of consumer websites and online definition resources in our dynamic hyperlinking algorithm. We believe that we are approaching the limits of the performance of an annotation method that works with just one term at a time without a deeper model of the clinical context.

Several systems reported in the literature have objectives similar to our own. A translator is implemented which automatically replace difficult terms with easier
synonyms in clinical text [4]. Infobuttons have been used to provide context-specific linking from online resources in the text of online Pap smear reports [3]. In addition, the health software development literature also contains some reports of using iterative and incremental development in software for consumer support [9][10].

A key limitation of the present study is that only a small number of experts were used to drive the iterative refinement; this cannot be expected to capture all the problems that health consumers might encounter and does not provide a quantitative measure of performance. An expert panel assessment of the correctness and helpfulness of SemLink’s dynamic hyperlinking approach for lay consumers has been undertaken [5]. We believe still further refinement is needed before the method is safe for a trial with health consumers (sufficiently free of misleading links), although one could query how safe it is for consumers to be left without support for their internet searches. Potential future applications include integration of SemLink with online personal health records or portals, and with patient online social networks.

5. Conclusion

The design process of SemLink involved iterative refinements to ensure that the semantic annotations and the associated hyperlinks selected by the system were appropriate to support health consumers in reading their discharge summaries. The system design was modified to reflect the identified issues. Findings from the iterative design sessions could help other designers of consumer informatics applications to better understand the natural language processing and dynamic hyperlinking problems that might be encountered by system which provide readability support in clinical text.

References