Aligning Lay and Specialized Passages in Comparable Medical Corpora

Louise DELEGER\textsuperscript{a} and Pierre ZWEIGNBAUM\textsuperscript{b}
\textsuperscript{a}INSERM UMRS 872 Eq. 20, Paris, France
\textsuperscript{b}LIMSI-CNRS, F-91403 Orsay, France

Abstract. While the public has increasingly access to medical information, specialized medical language is often difficult for non-experts to understand and there is a need to bridge the gap between specialized language and lay language. As a first step towards this end, we describe here a method to build a comparable corpus of expert and non-expert medical French documents and to identify similar text segments of lay and specialized language. Among the top 400 pairs of text segments retrieved with this method, 59\% were actually similar and 37\% were deemed exploitable for further processing. This is encouraging evidence for the target task of finding equivalent expressions between these two varieties of language.

Keywords. Natural Language Processing, Consumer Vocabulary, Comparable Corpora

Introduction

Over the years the general public has been dealing with an increasing amount of medical information, whether from the Internet or from gaining access to their own medical records as legislations allow them to. However this information is not always expressed in a language suitable to lay people. Specialized medical language is indeed difficult for non-experts and patients may not understand the information that is presented to them [1,2]. There is therefore a real need to empower patients with means which facilitate their understanding of medical language [3,4].

Previous work has addressed the issue in several ways. [5] developed methods to help lay people query the Web for health information. A number of papers emphasized the need for consumer health vocabularies [6,7] and investigated methods to this end [8]. [9] identified and defined difficult medical terms. Building a lexicon of linked specialized medical terms and lay expressions was also investigated as a way to bridge the gap between the two types of languages [10]. A preliminary step to many applications is to examine characteristics of medical texts [11,12]. [13] used such characteristics to automatically categorize expert vs. non-expert Web pages in Russian and French.
We aim at characterizing patient-friendly documents as opposed to expert documents in order to help design ways to adapt specialized medical information to patients. To do so we propose to build a comparable corpus of lay and specialized medical documents and to study the relations between the two sides of this corpus. More specifically, we would like to find correspondences between the two varieties of languages—lay and specialized—in these documents, in the same line as [10] but for French. This raises two issues: how to collect such a corpus (Section 1.1) and how to process it to identify corresponding passages (Section 1.2) in a setting where, as we shall see below, the comparability of documents is less ideal than in [10]. We then expose the evaluation conducted (Section 1.3) and its results (Section 2), discuss them and conclude (Section 3).

1. Material and Methods

1.1. Acquisition of a Comparable Corpus of Specialized and Lay Texts

We chose to work with medical documents dealing with the topic of tobacco and nicotine addiction and built a corpus of such documents in French, containing texts intended for experts and texts intended for the general public. Comparable corpora usually refer to multilingual text collections that address the same topic without being translations of each other. Here the notion is applied to monolingual texts from different genres—lay and specialized—but dealing with the same topic.

Building a corpus is far from being a trivial or an immediate task. A corpus must suit the needs and objectives of the target task, in terms of both size and quality. Today, a popular way of acquiring a corpus is collecting it from the Web [14], as it provides easy access to a virtually unlimited number of documents. When dealing with a Web corpus several issues arise. The first issue is the relevance of the retrieved documents to the targeted domain and is highly dependent on the method used to gather the documents. Several methods can be used for collecting a corpus from the Web on a specific topic: (1) general-purpose search engines, such as Google, can be queried with selected key words [15]; (2) domain-specific search engines can be used (in domains where they exist); (3) direct retrieval of relevant Web pages is also possible when these are known to the user (either directly or through a quick glance at the output of a Web search).

Another important issue specific to our type of corpus is the relevance to the targeted genre. We are gathering a comparable corpus of two genres—lay vs. specialized. Hence the need to classify each collected document as belonging to one genre or the other. Two main approaches are possible: either (a) relying on automatic text categorisation methods [16] (this type of approach is automatic but may introduce potential noise caused by categorization errors); or (b) using resources that include a classification of the documents or resources known to belong to one particular category.

We used methods belonging to points (2), (3) and (b). We discarded generic search engines since the documents would not be categorised and relevance to the domain would be especially questionable with a wide-range domain such as tobacco. We decided in favour of a more restricted search involving a small part of manual work.
First we queried two health search engines that we knew of (the health web portals CISMeF\(^2\) and HON\(^3\)) with a list of key words. Both provide access to trustworthy Web pages and allow the user to search for documents targeted to a population (e.g., patient-oriented documents). We also knew of relevant websites. Those were French governmental websites, including that of the HAS\(^4\) which issues guidelines for health professionals, and that of the INPES\(^5\) which provides educational material for the general public; as well as health websites dedicated to the general public, including Doctissimo\(^6\). The previous search results also allowed us to identify two more websites dealing with nicotine addiction and aiming at the general public: Tabac Info Service\(^7\) and Stop-tabac\(^8\).

Once collected, a corpus needs to be cleaned and converted into an appropriate format for further processing—i.e., extracting the textual content. HTML documents (as we have in our corpus) contain a certain amount of irrelevant information such as navigation bars, footers and advertisements—referred to as “boilerplate”—which can generate noise. While getting rid of HTML code is easy, identifying boilerplate is much more challenging \[17\]. Boilerplate removal methods can rely on HTML structure, on visual features (placement and size of blocks) and on plain text. We made use of the HTML structure in several ways. First, we noticed that some HTML documents contained meta-information as to the location of the content text in the pages—for instance, a block may have the attribute “class=content”, or comments such as “<!–content start-->” may be present. We then used HTML structure to identify potential navigation bars: we made the assumption that any list of links at the beginning of documents with a small amount of text had a high probability of being a navigation bar, so we removed such lists. Finally, we relied on the density of HTML tags to identify the start and end of content text in a document: we measured the ratio of text throughout the document in a sliding window of ten words. We defined the starting point for content extraction as the first location where the ratio became equal to 80% and the end as the last location where the ratio became equal to 80%. We also relied on plain text features to try to spot obviously irrelevant text, such as phone and fax numbers and e-mails (as often appear at the end of documents).

1.2. Identification of Relations between Specialized and Lay Texts

Our long-term objective is to find equivalent expressions in specialized and lay texts. As a first step, we tried to relate text passages taken from both sides of our comparable corpus which address similar topics and might thus contain corresponding expressions.

1.2.1. Topic Segmentation

Texts rarely constitute homogeneous units, as multiple topics are usually addressed in a single text. It might therefore be difficult to relate two entire texts, so we chose to work at the sub-level of topic segments. We performed topic segmentation on each text using the segmentation tool TextTiling \[18\]. A segment may correspond to one or several
paragraphs. Another possibility would have been to work at the paragraph level or even at the sentence level—this provides alternative ways to explore in further work.

1.2.2. Assessing the Similarity of Text Segments

Once our corpus was topically segmented, we tried to identify pairs of text segments addressing similar topics. For this we used common, vector-based measures of text similarity: Cosine and Jaccard. Both measures give a similarity score ranging from 0 to 1. We computed them for each pair of topic segments in the cross-product of both corpus sides. The input segments were transformed into vectors of words and their similarity was established based on the proximity of the vectors. Given two vectors $A$ and $B$, $\text{Jaccard}(A;B) = \frac{|A \cap B|}{|A \cup B|}$ and $\text{Cosine}(A;B) = \frac{A \cdot B}{\|A\| \|B\|}$. Cosine takes into account the frequency of the words while Jaccard only relies on the presence or absence of a word.

1.3. Evaluation

We evaluated the similarity results by selecting a sample (one every four) from the 600 pairs of text segments with the highest similarity scores, for each similarity measure. We obtained two samples of 150 pairs to review, one corresponding to the Cosine measure and one to the Jaccard measure. We evaluated them against the following two criteria: (1) whether the text segments were actually similar; (2) whether the text segments contained equivalent expressions, which a future step will aim at detecting automatically—i.e., whether they were exploitable for further processing.

2. Results

Table 1 shows the size of our corpus. Both sides (lay and specialized) of the corpus contain approximately the same number of words but the number of lay documents is far greater. This means that our search brought more lay documents, but also that expert texts are much longer. Topic segmentation resulted in 3,226 segments in specialized texts and 2,769 in lay texts (see table 2).

Evaluation results for the most similar segment pairs are given in figure 1. We see that the proportion of similar pairs and of exploitable pairs is rather good up to the 400th pair (59% and 37% with Cosine) which slowly decreases afterwards, for both similarity measures. This shows that we may draw the limit of similar pairs to the first 400 ones, and discard the other pairs. The proportion is also almost always higher with Cosine which seems to indicate that this measure is more appropriate. A prospect for assessing the similarity would be to use a combination of those two measures as a similarity score, instead of just one or the other.

<table>
<thead>
<tr>
<th>Documents</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specialized</td>
<td>61</td>
</tr>
<tr>
<td>Patient-friendly</td>
<td>546</td>
</tr>
<tr>
<td>Total</td>
<td>607</td>
</tr>
</tbody>
</table>
Table 2. Number of segments produced by topic segmentation

<table>
<thead>
<tr>
<th></th>
<th>Specialized</th>
<th>Patient-friendly</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nb of segments</td>
<td>3,226</td>
<td>2,769</td>
<td>5,995</td>
</tr>
</tbody>
</table>

Figure 1. Results for the 600 best segment pairs

Table 3. Excerpt from a pair of text segments

<table>
<thead>
<tr>
<th>Specialized</th>
<th>Lay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le tabagisme de l’homme serait à l’origine :</td>
<td>Les effets délétères du tabac sur la fertilité en général :</td>
</tr>
<tr>
<td>1. d’une dysfonction érectile réversible à l’arrêt ; . . .</td>
<td>Chez l’homme : Dysfonction érectile (qui disparaît à l’arrêt du tabac) ; . . .</td>
</tr>
</tbody>
</table>

An excerpt of similar text segments is given in table 3 (equivalent expressions are highlighted in corresponding colors). Similar but non-exploitable pairs include navigation bars that could not be filtered out during conversion from HTML to text (hence the need to clean the corpus as much as possible after collecting it), and portions of texts describing statistics.

3. Discussion and Conclusion

We built a comparable corpus of lay and specialized medical documents in French from which we were able to identify pairs of similar text segments that seemed exploitable for finding equivalent expressions between lay and specialized language. This is encouraging evidence since these documents, although addressing the same general topic of tobacco and nicotine addiction, were a priori rather different since they come from various sources and are targeted to different populations.

The number of such pairs of similar segments may seem rather small compared to all potential segment pairs. However, as above mentioned, we are dealing with a comparable corpus of a priori rather dissimilar documents and not with a parallel
corpus, in contrast to [10] whose corpus had paired lay and specialized texts. This also raises the issue of the corpus size needed to provide sufficient coverage of the domain. The limited number of selected text segments may indicate that we need to build a larger corpus.

Future work includes finding equivalent expressions (sentences, phrases, words) among the pairs of text segments; testing a different similarity measure that would be a combination of the two measures used in this work; segmenting the texts into paragraphs or into sentences instead of topic segments.

Finally we can wonder whether results would be different with a different topic than tobacco. Testing the method on a different corpus is also a prospect of this work.

To sum up, we described a method to build a corpus of expert and non-expert medical French documents and identify similar text segments of lay and specialized language, as a first step towards bridging the gap between these two varieties of language. We were able to identify similar text segments. Although there is room for improvement and further testing, this gives good hope for our target task of finding equivalent expressions between lay and specialized medical language.

References