The Contribution of Morphological Knowledge to French MeSH Mapping for Information Retrieval

P. Zweigenbaum, Ph.D.,¹ S.J. Darmoni, M.D., Ph.D.,² N. Grabar, M.Sc.¹

¹ DIAM — Service d’Informatique Médicale, DSI, Assistance Publique – Paris Hospitals & Département de Biomathématiques, Université Paris 6, Paris, France
pz, ngr}@biomath.jussieu.fr http://www.biomath.jussieu.fr/

² Computer and networks department, Rouen University Hospital & Perception, Information and Systems Lab, National Institute of Applied Sciences, Rouen, France
stefan.darmoni@chu-rouen.fr

MeSH-indexed Internet health directories must provide a mapping from natural language queries to MeSH terms so that both health professionals and the general public can query their contents. We describe here the design of lexical knowledge bases for mapping French expressions to MeSH terms, and the initial evaluation of their contribution to Doc’CISMeF, the search tool of a MeSH-indexed directory of French-language medical Internet resources. The observed trend is in favor of the use of morphological knowledge as a moderate (~5%) but effective factor for improving query to term mapping capabilities.

INTRODUCTION

The Internet has become a major source of health information for the health professional and the citizen. However, general directories and search engines do not permit the end-user to obtain a clear and organized range of available useful health information. Therefore, a number of specialized health directories and catalogs have been created, such as CliniWeb, HON and CISMeF, some of which (including those three) are additionally indexed with the MeSH Thesaurus and provided with a search interface. This allows complex queries to be stated, which can take advantage of the hierarchical structure of the MeSH (e.g., ‘explode’ function in Doc’CISMeF). Since not all users know the MeSH terms, the search interface must provide some way of mapping so-called “Natural Language” queries to MeSH terms. Mapping techniques try to reconcile character-level differences (typos, uppercase, accents) as well as morphological, syntactic or semantic variants. To our knowledge, no study has been done yet on the specific importance of coping with each of these variations. The present work addresses morphology. It studies the specific impact of the provision for morphological variants in French MeSH mapping on the access to the CISMeF Internet health information directory through its Doc’CISMeF search tool.

Several types of techniques have been designed to perform “approximate” term matching, i.e., to identify target terms that are close to, but not identical to, query terms. String-based techniques can cope with character-level differences such as typos, and can be implemented very efficiently. Stemming is an algorithmic technique for reducing a word to its “stem” by removing common affixes, so that words that belong to the same morphological family (e.g., {probability, probabilistic}) are considered identical for matching purposes. Linguistic knowledge can provide a more accurate account of morphological variation in inflection deals with the grammatical variation of a single word (number, gender, etc.), derivation adds affixes to a base word form to produce new words (e.g., infection, infectious) and compounding combines several radicals to obtain complex words (e.g., hypercalcemia). Parsers can deal with syntactic variation, e.g., inserting modifiers and conjunctions, or changing word order and part-of-speech. Semantic techniques often refer to the substitution of non-morphologically-related synonyms (e.g., {heart, cardiac}), or to the use of links to related notions such as hyponymy or meronymy (e.g., {myocardium, heart}), typically
drawn from structured terminologies. Statistical methods can help identify semantically related words, e.g., by exploiting word co-occurrence in large text collections, as done by the Ovid system (gateway.ovid.com). Concept-based matching either is another name for semantic techniques, or assumes a symbolic representation of information on which deeper matching is performed.

We focus here on morphology-based techniques for French. It has been shown that stemming techniques, while widely used for the English language in Information Retrieval (IR) tasks, are not sufficient for languages with a richer morphology such as French, hence the lack, to the authors’ knowledge, of a port of any of the common English stemmers to French. Linguistically-grounded morphological knowledge and techniques are therefore needed. Some work has been done on morphological techniques for French medical language processing. Researchers at Xerox Research Center Europe have evaluated the impact of morphological and syntactic techniques on an IR task on a French newspaper corpus. They showed that inflectional knowledge brings a significant increase in average precision, whereas derivational knowledge brings a non-significant additional improvement. However, on the one hand, no morphological knowledge base is yet publicly available for French; inflectional knowledge has been available for some time, but derivational knowledge is still a rarer resource. On the other hand, the contribution of morphological knowledge to medical IR remains to be assessed.

We present here such knowledge and techniques, and an experiment to evaluate their impact on an Information Retrieval subtask: matching natural language queries submitted to the Doc’CISMeF search tool to French MeSH terms. It relies on a log of queries to Doc’CISMeF, and aims to assess the differential contribution of morphological knowledge to term matching in this context.

**BACKGROUND AND MATERIAL**

Doc’CISMeF (D’C) is a generic search tool based on an information structure model which encapsulates the MeSH thesaurus. It was launched in June 2000, and has received since then a steadily increasing number of queries per day – in January 2001, an average of over 1,200 a day from 400 unique users. We extracted from the http server log all queries sent to the Doc’CISMeF search “servlet”. We discarded: (i) the first two months of operation, in order to eliminate potential startup effects; (ii) all queries sent from within the CISMeF team and more widely from the Rouen University Hospital, where users have received a specific training about the MeSH thesaurus (a monthly 2 hour training session), and from the Rouen INSA engineering school, which is linked to the CISMeF team; (iii) all queries performed through the “advanced” interface; and (iv) all empty queries. The number of occurrences of each query is irrelevant, since for technical reasons, one query may result in several lines in the log. Our corpus of queries totals 27,029 queries; among these, the August queries (2,389) were used to debug the system.

The target terms are those indexing Doc’CISMeF: the French MeSH (19,971 terms and 83 qualifiers), augmented with 38 metaterms and 101 resource types, including some accented variants. Morphological knowledge was automatically derived from SNOMED and ICD-10 in previous work. It includes: (i) pairs of morphologically related words (e.g., \{abdomen, abdominal\}, \{abdominal, abdominale\}); (ii) morphological rules (e.g., enjinal); (iii) words and rules tagged with part-of-speech information (e.g., \{muscle/NN, musculaire/ADJ\}); (iv) pairs and rules with “lemmatized” words: each word is replaced with its uninflected form (e.g., singular masculine for adjectives; this is the case for the tagged pair above).

**METHODS**

**Design of the Lexical Knowledge Bases**

We collected from the above data (lemmatized, tagged word forms) a set of pairs \{lemma, inflected form\}, from which we further derived and manually completed inflection paradigms for the words in ICD-10 and SNOMED. 2906 unique inflection pairs, corresponding to 1224 families.
(e.g., apical, apicale, apicaux, apicales) and 4125 different word forms, were collected. We shall consider that, for term matching purposes, any of the forms in a family is equivalent to the others. For derivational knowledge, we started from the tagged and lemmatized word pairs. Derivation pairs are generally substitutable for Information Retrieval. Compound pairs are more complex to use, and have been reserved for further investigation. To separate compounds from derivations, we used the following heuristic: most pairs with two different parts of speech are derivations (e.g., from noun to adjective), whereas most pairs with identical parts of speech are compounds (\{lymph/NN, lymphoblaste/NN\}). This rough division was then adjusted manually. We finally collected 1042 derivation pairs (794 distinct families, e.g., aorte, aortique, aorto) for 1759 lemmas. When merged, inflection and derivation knowledge involve 1600 families and 5462 word forms (e.g., immun, immune, immunes, immunisation, immunité, immuno, immuns). Note that these lexical knowledge bases were not specifically prepared for the Doc’CISMeF (MeSH) vocabulary. This will be performed in a forthcoming experiment.

Approximate Term Matching
A query is first segmented into words according to whitespace and punctuation, which are filtered out. All words are transformed into lower case. This results in a set of word forms, from which stop words are removed (\{order\} method). The remaining are considered “content” words. The next steps (“query expansion”) add “equivalent” word forms to each content word if appropriate. These word forms may include reaccentuated or disaccentuated forms if they exist in the target vocabulary (\{accent\}); and other inflected forms (\{infl\}) or derived words (\{deriv\}) depending on the morphological knowledge provided. The result for each input query word is a disjunction of “equivalent” word forms (e.g., muscle/musculaire/musculaires); and the result for a query is a conjunction of such disjunctions (e.g., (personnes/personne) AND (agees/age/agee)).

Each target term is segmented and lower-cased as a query, but not further processed. It is then handled as a “set” of words – i.e., word order and repetition are not significant. Given a query, target terms are ranked first if they (in the specified order): (i) satisfy the largest number of disjunctions (contain the largest number of query words); (ii) have the smallest number of extra words (extra); (iii) contain the largest number of exact words forms from the original query (resort less to “equivalents”); (iv) contain words closer to the beginning of the query; in case of a tie, the final decision is alphabetic order. A “greedy” algorithm is used to successively select target terms in order to “cover” the words of the query. In addition, we also tested a stricter version of clause (ii) of this algorithm (\{noextra\}) where answer terms may only contain words present in the (expanded) query (plus possibly stop words).

The algorithm was implemented within an existing term matching program written in Perl5; matching speed is reasonable for testing purposes at about 600 queries/mn on an HP-UX machine. Called as a CGI program, it returns an HTML page which presents the user with queries that s/he can send to Doc’CISMeF in a click (www.biomath.jussieu.fr/cismef/).

Evaluation
Our MeSH mapping task can be characterized as finding the conjunction of MeSH terms that together best characterize the user’s information need. Since the output of a query is one conjunction of terms, the traditional IR measurements of recall and precision are ill suited. Instead, we measured on the one hand the overall impact of
Table 1: Overall impact of morphology.

<table>
<thead>
<tr>
<th>variant</th>
<th>exact MeSH term</th>
<th>changed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exact order accent</td>
<td>infl deriv</td>
</tr>
<tr>
<td>extra</td>
<td>182 1364 1679</td>
<td>429 128</td>
</tr>
<tr>
<td>noextra</td>
<td>256 106</td>
<td></td>
</tr>
</tbody>
</table>

mapping parameters: the number of queries for which a change in the mapping method (order, accent, infl, deriv) causes a change in the result. On the other hand, we evaluated the quality of these changes, i.e., whether the changed answers are better or worse than the initial ones. Answer quality was rated by the CISMeF team (one medical informatician and two medical librarians, by consensus) on a 4-value scale, from 0 (very bad) to 3 (very good), according to their relevance for querying CISMeF contents. As a baseline, they selected a set of 58 queries with two words or more, that were not exact target terms and had no spelling errors, and rated their answers (accent, extra) to assess the difficulty of the mapping task. In a complementary experiment noextra, we checked whether from a linguistic point of view the changed answers correctly added or refined the MeSH terms of the initial answer, or added an erroneous term. The matching algorithm was run on all queries of each month in the Doc’CISMeF log corpus. The general pattern observed was the same for all months, so we took Sep. 2000 as a representative for closer examination (6469 different queries).

RESULTS

Table 1 shows the overall impact of mapping parameters on answers to queries (month 09/2000). The first three columns (exact MeSH term) show the number of queries that were actually one exact MeSH term (2.8%), modulo word order and stop words (21.1%) and accents (26.0%). The next two columns display the number of queries whose answers changed when adding inflectional knowledge to [accent] (6.6%) and then derivational knowledge (2.0%). The last row concerns the [noextra] variant (4.0% and 1.6%).

The baseline evaluation of non-morphological mapping quality (accent, extra) obtained an average 1.72 grade ($\sigma = 1.18$, max=3). For the [extra] strategy, a sample of 20 answers out of the 429 modified by inflection was examined: their average grade increased from 0.53 to 1.07. Half the [deriv]-changed answers were examined, and went up from 1.14 to 1.91. For the [noextra] strategy, we examined half the [infl]-changed answers and all the [deriv]-changed ones. Figure 1 shows the results: over 83% of the changed answers are improved from a linguistic point of view. Table 2 contains examples of improved behavior.

DISCUSSION AND CONCLUSION

This experiment mainly involves a single technique: morphology-based query expansion. It is meant as a methodological test for that technique in isolation. It is only one block in a series of complementary techniques, currently in construction, that are to work in cascade to match queries to target terms. Its implementation can further be improved by expanding and adapting our morphological knowledge over the MeSH and by training the mapping algorithm on the log queries.

Query expansion in [infl] and [deriv], by nature, increases the set of potential answer terms. The [noextra] decision variant aims to prune this set to answer terms that are lexically more focussed on the query. It removes spurious mappings such as grands brûlés $\rightarrow$ grands singes, maladies (deeply burnt $\rightarrow$ ape diseases). This additional filtering makes the contribution of morphological knowledge clearer in that setting: it allows 4.7% of the queries to be better mapped, while adding a small amount of noise (0.9%).
Table 2: Example contributions of morphological knowledge (the slash ‘/’ separates answer terms).

<table>
<thead>
<tr>
<th>query</th>
<th>answers [accent]: [extra]</th>
<th>answers [inf]: [deriv]: [noextra]</th>
</tr>
</thead>
<tbody>
<tr>
<td>aberrations chromosomiques/anomalies</td>
<td>aberrations chromosomiques</td>
<td>aberration chromosomique, anomalies</td>
</tr>
<tr>
<td>hematome pelvien</td>
<td>hematome / membre pelvien</td>
<td>hematome / pelvix</td>
</tr>
<tr>
<td>tumeur du glomus</td>
<td>glomus carotidien, tumeur</td>
<td>glomique, tumeur</td>
</tr>
<tr>
<td>kyste du rein</td>
<td>rein / kyste arachnoide</td>
<td>rein kystique</td>
</tr>
<tr>
<td>stenose valve aortique</td>
<td>stenose isthmique aorte congenitale / prolapus valve aortique</td>
<td>stenose aortique valvulaire</td>
</tr>
<tr>
<td>tumeur bronchique</td>
<td>face, tumeur / fistule bronchique</td>
<td>tumeur bronche</td>
</tr>
</tbody>
</table>

These experiments confirm what was observed by the Xerox team[10] with here a moderate but positive impact of both inflection and derivation on query results. This is an encouragement to continue to build more complete morphological resources for medical language to further increase the impact of morphological query expansion.

The experiments performed illustrate how the task specification and the specific answer selection criteria can change the appreciation of the differential contribution of morphological knowledge in IR. A global, IR-oriented evaluation is therefore necessary, assessing the impact of these factors on the final retrieval of documents (indexed Web sites). It is currently under way.

References


