The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

Matching controlled vocabulary words

Natalia Grabar[a], Pierre Zweigenbaum[a], Lina Soualmia[b,c], Stéfan Darmoni[b,c]

[a]*STIM/DSI, Assistance Publique – Hôpitaux de Paris, France*
[b]*L@STICS, Centre Hospitalier Universitaire de Rouen, France*
[c]*Perception, Information and Systems Lab, National Institute of Applied Sciences, Rouen, France*

Abstract

This study examines an enabling condition for natural language access to medical knowledge resources (Medline, CISMeF) indexed with controlled vocabularies (e.g., the MeSH): is the vocabulary of user queries comparable with that of the index terms? The two vocabularies were compared in their original form, then under incrementally normalized forms, using character-based normalizations then linguistic normalizations. Only 16.7% of the user vocabulary, in its original form, is in the MeSH. Progressive normalizations increase this proportion to 65.5%. Besides, if the frequencies of occurrence of words are taken into account, 89.3 % of user word occurrences can be matched to MeSH words. This shows the interest of taking into account further matching methods between queries and index terms than those presented here.

Keywords:
Controlled Vocabulary; Natural Language Processing; World Wide Web; Language; France

1 Introduction

A growing quantity of medical knowledge is available to health care professionals and to the general public: bibliographic databases like Medline and various kinds of documents (clinical guidelines, thematic sites, general gateways). To help the knowledge 'consumers' find these resources, a possible approach consists in *indexing* the documents with a *controlled vocabulary*. For instance, Medline (www.ncbi.nlm.nih.gov/pubmed) and Internet catalogs like CliniWeb (www.ohsu.edu/cliniweb), CISMeF (www.chu-rouen.fr/cismef) or HON (www.hon.ch) index the information they contain with MeSH terms (www.nlm.nih.gov/mesh). One of the main reasons to use a thesaurus like the MeSH is its hierarchical structure, which allows to query about a generic notion (e.g., *"cardiac diseases"*) and to retrieve all the 'documents' indexed with more specific terms (e.g., *"myocardial ischemia"*). It is then advantageous to use these pivot terms to access medical knowledge. However, one cannot expect all users to know every term in a controlled vocabulary. The above sites allow a user to formulate natural language queries and find the best match with indexing terms. We tackled this problem in previous work [1], studying the differential contribution of linguistic knowledge (inflection and derivation) in the matching task. The question we address now involves a precondition for this mapping: to which extent is the vocabulary of the submitted free queries comparable with the target vocabulary?

We first present previous work on matching natural language expressions to controlled vocabulary terms. We specify the queries and the target terminology on which we worked. We explain the methods designed to match and normalize query vocabulary and target vocabulary. We then detail the results of these comparisons, discuss their implications and conclude on recommendations for further work.

2 Background

Many works have addressed 'terminological variation', which can be processed at different levels: characters [2] (spelling and accenting mistakes, case variants), words and their morphological variants [3, 4, 5], syntax [5], concepts with general-language [6] or domain-specific [7] synonyms, or through the study of the distributional similarity of words in text corpora [8]. Phonemic matching is yet another method to match words based on their pronunciations. Finally, recent works [1, 9, 10] show the contribution of morphological processing for information retrieval in French. The general observation is that lemmatization brings a statistically significant improvement and that stemming additionally improves the results, but in a non-statistically significant way.

3 Material used

The *queries* studied are those received by the 'simple search' interface of Doc'CISMeF [11] (doccismef.chu-rouen.fr), from September 2000 to January 2001. We filtered out from the server log empty queries and queries issued by the CISMeF team. This yields 108,660 queries (29,092 unique queries), or 76,341 occurrences of unique 'query.machine' pairs. The *target terms* indexing CISMeF come from the French version of the MeSH [12] (19,971 terms, 9,151 synonyms and 83 qualifiers), enriched with CISMeF's 38 'metaterms' and 101 'resource types' [11]. This results in 29,035 different terms (mixed case, mixed accentuation). *Morphological knowledge* was reused from previous work [1] and extended. We use it to *lemmatize* (*e.g.*, French *"abdominaux"* → *"abdominal"*) and to *stem* words (*"abdominal"* → *"abdomen"*). We used 308,847 pairs for lemmatization (compiled from general dictionnaries and medical corpora: MENELAS, CLEF) and 1,041 pairs for stemming [1, 13]. Moreover, general morphological rules (cutting of final *-s*) are used. Our *stop word* list contains 344 words.

4 Comparing vocabularies

The method consists in tokenizing source (queries) and target (MeSH) terms into words and then in matching the resulting vocabularies after successive normalizations. Tokenization is done on all non-alphanumeric characters and gives a set of source words and a set of target words. These sets allow a first evaluation of vocabulary overlap. The next treatment is performed on source subset which is not yet matched with the target vocabulary and submitted to the two types of normalizations (character- and linguistic-level). Stop words are eliminated from both vocabularies at an early stage.

Character normalizations We apply two types of normalization at this step. (1) *Lowercase conversion*: all the uppercased caracters are replaced by their lowercase version; (2) *Deaccenting*: all accented caracters (*e.g.*, *"èéêë"*) are replaced by non-accented (*"e"*). This kind of processing is useful because words in the French MeSH are not accented, and words in queries can be accented or not, or wrongly accented (*"athlètisme"*).

Morpholexical normalizations This type of normalization uses linguistic knowledge. The *lemmatisation* reduces an inflected form to its canonical form. We applied three types of lemmatizations: (*i*) a list of 308,812 {*"lemma"*, *"form"*} pairs; (*ii*) a heuristic for deleting regular plurals (which obtains good results in information retrieval [10]) mainly deletes final -

s, substitutes *-aux* with *-al* and deletes final *-x*; (*iii*) the combination of both. Stemming is then applied with 1,041 {*"base"*, *"form"*} pairs. 'Stemming' reduces a dedived form to its base.

Spelling correction Words in the remaining subsets might be misspelled words. We decide to attempt *spelling correction* on these words using the Unix `ispell` tool, with the target vocabulary as reference dictionary.

Types and occurrences At each stage of processing the number of unique words (*types*) and the number of *occurrences* of each word have been computed: occurrences take into account the frequence of each word type. The total number of occurrences remains constant along successive normalizations, except when stop words are removed. In contrast, the total number of word types decreases as some types are replaced with a normalized form.

5 Results obtained

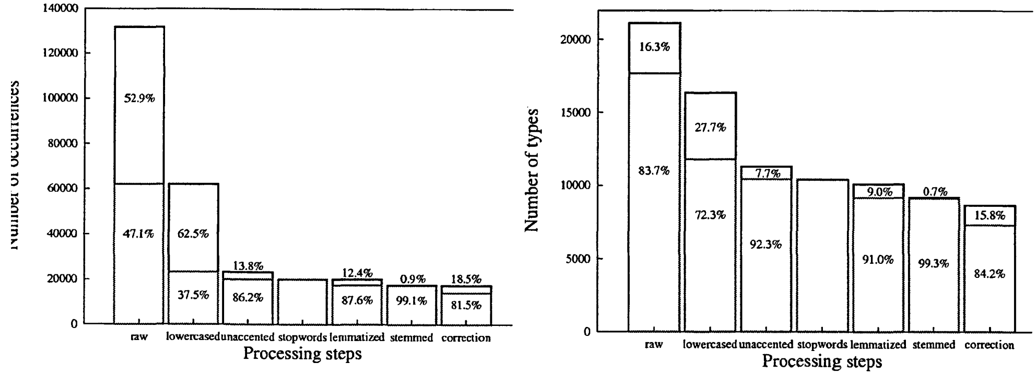

Figure 1: Evolution of occurrence and type matches at each step.

Figure 1 details results of occurrences (left) and types (right) matching for the global period from September 2000 to January 2001. Queries contain 29,092 different terms, corresponding to 21,112 different unique words (types) and to 131,570 occurrences (box "row", y axis). Among these words, 3,438 types (69,602 occurrences) are common with the MeSH (21,475 types and 58,912 occurrences); and 17,674 types (61,968 occurrences) are unknown: only 16.3% of types and 52.9% of occurrences are matched at this step. These percentages are displayed in box "raw"; the upper part shows the maches, and the lower part the unmatched words. Lowercase conversion allows to recognize 62.5% more occurrences (27.7% types) of the remaining sets (box "lowercased"). 23,217 occurrences (11,806 types) remain still unknown. Deaccenting (box "unaccented") allows to match 13.8% more occurrences (7.7% of types). At this step 20,004 occurrences (10,420 types) remain unknown. Eliminating 33 stop words corresponds to deleting 85 occurrences. Indeed, this produces no new matches. Lemmatization (box "lemmatized") matches 12.4% more occurrences (9% types), and stemming (box "stemmed") about 1% occurrences (0.7% types). A manual study of words remaining unmatched before spelling correction gives the following typology: misspelled words (*"acetylsalycilique"*, *"altzheimer"*, *"1lzheimer"*); morphological variants of MeSH words, for which our morphological resources are uncomplete (*"cicatriciel"* would have matched *"cicatrice"* in MeSH, *"encephalique"* would have matched *"encephale"* in MeSH);

abbreviations, (*"bpco"*, *"esb"*, *"biam"*); English words, (*"allergy"*, *"bronchiolitis"*); concepts not in MeSH, (*"calcemie"*, *"adenomectomie"*), including proper nouns (*"beux"*, *"broussais"*). Spelling correction (box "correction") proposes to correct about 15.8% of remaining types (18.5% of occurrences). Finally, 65.5% of types are recognized (89.3% of occurrences) of the initial set.

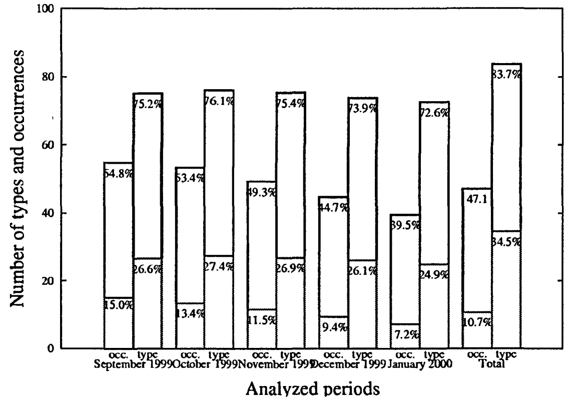

Figure 2: Monthly evolution of unmatched occurrences and types (Sept. 2000 to Jan. 2001).

Figure 2 displays the monthly evolution of unmatched words for the same period: both occurrences (occ.) and types (type) are indicated in their raw (higher boxes) and fully normalized forms (lower boxes). Figures are given for raw matching (higher box; initial raw step of figure 1 and for fully normalized matching (lower boxes; "correction" step of figure 1). It shows that the percentage of unknown words decreases both in occurrences (steeply) and in types (more slowly). We indicate the global result (box "Total"): when all the received queries are processed together. Matching of occurrences corresponds to the average result as for the monthly evolution. In contrast, the number of unmatched types increases: new unknown types show up each month.

6 Discussion

The comparison of the original word sets shows that about one half of source occurrences and only one sixth of types match target CISMeF MeSH words. This is a low proportion, which means that without normalization, about one half of user queries would obtain lower-quality results. As a matter of fact, in Doc'CISMeF, queries with terms outside the MeSH generate a full text search on Doc'CISMeF records. We have seen that this first result was significantly improved by normalization treatments. If we take into account, though, that users write mixed case and mixed accentuation queries, this initial proportion of matches can be considered as high. Normalizations allow to recognize up to 89.3% of occurrences (65.5% of types).

The analyses of monthly evolution of matching shows an improvement for occurrences and only slow progress for types. This improvement can be explained by: (1) an increase in the number of 'direct' links to Doc'CISMeF proposed by some sites; (2) an extension of CISMeF MeSH with common pathologies (written in lowercase); (3) an important number of

Doc'CISMeF users being librarians and acquainted with MeSH (... in English); (4) and a self-training of users to CISMeF contents.

Spelling correction remains an important factor for additional matches. Other analyses of query logs [14] show a similar pattern to that observed here for unknown words: 13% of queries included a spelling error. However, because the result of spelling correction is too uncertain, in this experiment we did not run it on short (less than 5 letters) words, and only nonambiguous corrections were taken into account (one possible correction by word). To ensure more reliable results, we must now enforce more constraints on spelling correction and evaluate its results.

The largest improvement is obtained by character-level normalizations (relative improvement of 27.7% on types and 62.5% on occurrences). The global contribution of morphological knowledge is lower (relative improvement in the order of 9% of types and 13% of occurrences), but requires more initial resources. It reaches 56.8% of types and 86.8% of occurrences of the user vocabulary. Here as in many problem-solving situations, beyond a certain point, the additional effort needed grows while the corresponding improvements decrease. Besides, the contribution of morphological processing is higher in terms of types than occurrences; it is therefore useful to allow users to submit a larger variety of queries. The coverage of morphological resources (lemmatizing as stemming) is not exhaustive. Lemmatizing resources can be enriched through existing tools for syntactical processing (tagging, lemmatizing) and the human validation. While stemming resources are missing for French [15]. The UMLF project [16] should allow to complete them substantially.

The same processing should now be applied to other, more recent months of the CISMeF log to confirm these observations. Finally, words that cannot be matched with the MeSH vocabulary might still occur in the CISMeF records. If this case, the classical model of information retrieval could be applied there as a backoff model.

7 Conclusion

The evaluation presented in this paper shows the importance of low-level processing to match user and indexing vocabularies: they are now implemented in Doc'CISMeF. It also shows the relevance and limits of morphological knowledge in this task. Spelling correction is an important factor of improvement, but its results are less reliable and its use must be constrained. The proportion of remaining unknown words points out the necessity to apply other methods, such as using synonyms [7] or distributional similarities, or to backoff to full-text search. The results of this study provide a useful measure of the adequacy of users and indexing vocabularies in a natural language query interface to a Health gateway. They should become one of the metrics used for the routine follow-up of Doc'CISMeF.

8 References

[1] Zweigenbaum P, Darmoni SJ, and Grabar N. The contribution of morphological knowledge to French MeSH mapping for information retrieval. *J Am Med Inform Assoc* 2001;8(suppl):796–800.

[2] Lovis C and Baud R. Fast exact string pattern-matching algorithms adapted to the characteristics of the medical language. *J Am Med Inform Assoc* 2000;7(4):378–91.

[3] McCray AT, Srinivasan S, and Browne AC. Lexical methods for managing variation in biomedical terminologies. In: Proc Eighteenth Annu Symp Comput Appl Med Care, Washington. Mc Graw Hill, 1994:235–9.

[4] Lovis C, Michel PA, Baud R, and Scherrer JR. Word segmentation processing: a way to exponentially extend medical dictionaries. In: Greenes RA, Peterson HE, and Protti DJ, eds, Proc 8^{th} World Congress on Medical Informatics, 1995:28–32.

[5] Jacquemin C and Tzoukermann E. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In: Strzalkowski T, ed, *Natural Language Processing and Information Retrieval.* Kluwer, Boston, MA, 1999:25–74.

[6] Hamon T, Nazarenko A, and Gros C. A step towards the detection of semantic variants of terms in technical documents. In: Proc 17th International Conference on Computational Linguistics (COLING-ACL'98), Montréal, Quebec, Canada. 1998:498–504.

[7] Pouliquen B, Delamarre D, and Le Beux P. Indexation de textes médicaux par extraction de concepts, et ses utilisations. In: Actes de JADT, 2002:617–28.

[8] Xu J and Croft BW. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems* 1998;16(1):61–81.

[9] Gaussier E, Grefenstette G, Hull D, and Roux C. Recherche d'information en français et traitement automatique des langues. *TAL* 2000;41(2):473–93.

[10]Savoy J. Morphologie et recherche d'information. Cahier de recherche en informatique CR-I-2002-01, Université de Neuchatel, Division économique et sociale, Faculté de Droit et des Sciences Économiques, 2002.

[11]Darmoni SJ, Thirion B, Leroy JP, et al. A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the Internet. *Med Inform Internet Med* 2001;26(3):165–78.

[12]Institut National de la Santé et de la Recherche Médicale, Paris. Thésaurus Biomédical Français/Anglais, 2000.

[13]Grabar N and Zweigenbaum P. A general method for sifting linguistic knowledge from structured terminologies. *J Am Med Inform Assoc* 2000;7(suppl):310–4.

[14]David Hawking PB and Craswell N. An intranet reality check for TREC ad hoc. Technical report, CSIRO Mathematical and Information Sciences, Canberra, Australia, 2000. Available at http://pigfish.vic.cmis.csiro.au/ nickc/pubs/.

[15]Hathout N, Namer F, and Dal G. An experimental constructional database: the MorTAL project. In: Boucher P, ed, *Many morphologies.* Cascadilla Press, Somerville, MA, 2002:178:209.

[16]Zweigenbaum P, Baud R, Burgun A, et al. Towards a Unified Medical Lexicon for French. In: Medical Informatics in Europe, 2003. Submitted to MIE 2003.

Address for correspondence

Natalia Grabar, Mission de Recherche en Sciences et Technologies de l'Information Médicale (STIM),
DSI, Assistance Publique – Hôpitaux de Paris, 91, boulevard de l'Hôpital, 75634 Paris Cedex 13, France
E-mail : ngr@biomath.jussieu.fr Url : http://www.biomath.jussieu.fr/~ngr/