Extracting Linguistic Knowledge from an International Classification

Robert Baud [1], PhD, Christian Lovis [2], MD, Anne-Marie Rassinoux [3], PhD, Pierre-André Michel [1], Jean-Raoul Scherrer [1], MD

[1] *Division d'Informatique Médicale and* [2] *Département de Médecine.*

[3] *Division of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA*

University Hospital of Geneva, 1211 Geneva 14, Switzerland

email: Robert.Baud@expasy2.hcuge.ch

Automatic extraction of knowledge from large corpus of texts is an essential step toward linguistic knowledge acquisition in the medical domain. The current situation shows a lack of computer-readable large medical lexicons, with a partial exception for the English language. Moreover, multilingual lexicons with versatility for multiple languages applications are far from reach as long as only manual extraction is considered. Computer-assisted linguistic knowledge acquisition is a must.

A multilingual lexicon differs from a monolingual one by the necessity to bridge the words in different languages. A kind of interlingua has to be built under the form of concepts to which the specific entries are attached. In the present approach, the authors have developed an intelligent rule-based tool in order to focus on a multilingual source of medical knowledge like the International Classification of Disease (ICD) which contains a vocabulary of some 20'000 words, translated in numerous languages.

The ICD-10 classification

This paper describes the process of extracting lexical knowledge from a multilingual classification like ICD10 as undertaken in the authors group. This is an important step because it has been possible, with limited manpower resources, to semi-automatically extract a substantial amount of knowledge from a multilingual corpora. This kind of reverse engineering on an existing classification has been surprising by the ease to perform it and the richness of the harwest. This successful experience could possibly be repeated with other corpora of medical texts.

The goal : A Multilingual Lexicon

The availability of large medical lexicons in a multilingual context is not today a reality. They are numerous lexicons available here and there, but they are essentially monolingual. The UMLS Specialist lexicon [i] is quite large with more than 60'000 entries but it is not really multilingual and there is a lack of publications about its usage in another language than English. This is true despite this lexicon has a large number of French entries, because they are given only in upper case letters and therefore they are practically not usable without the accents! It has also to be mentioned the existence of a 9 languages medical lexicon of nearly 2000 words available on Internet[ii]. Nevertheless, such an effort is practically not significant by its size which is definitely to small for practical applications. This is true despite this project gave considerable attention to the validation process. This situation raises the following question : Why multilingual lexicons does not exist or are not made available to the scientific community?

The explanation lays certainly in the fact that there is an important design problem with multilingual lexicons. A multilingual lexicon is not just the juxtaposition of multiple monolingual lexicons. A multilingual lexicon necessitates a link between corresponding words in different languages and the task of defining these links may be tremendous. The point of convergence from the corresponding words in different languages leads to the notion of concept : words in different languages are different signs (or pointers) to a unique concept. This problem has been described under the key idea of the knowledge triangle as applied to medical linguistic[iii]. Language independent concepts are the focus point necessary for multilingual lexicons. This structure is inherent to multilingual lexicons and it has been described by the authors elsewhere[iv].

When accepting the notion of concepts as an interlingua between different languages, and when aiming at the construction of lexicons with some 20'000 entries, one is faced to the problem of organising the concepts. Any structuring action about concepts is clearly the beginning of a model of the domain and such a task requires heavy resources. This is the main reason why true multilingual lexicons are not yet been developed in the medical domain.

Why Automatic Extraction ?

Automatic extraction of lexical knowledge from any source has different advantages to underline now. There are at least three aspects to be mentioned :

- Lexicon availability is the major bottleneck for NLP tools today and this is due to the lack of available manpower resource. An automatic extraction scheme is supposed to save manpower. This is basically true when dealing with five or more languages.
- An in-depth knowledge of all the languages is not strictly necessary. Indeed, only a subset of surface syntactic knowledge is required at least in a first step of extracting the relevant vocabulary and it may be designed and implemented once for all. Nevertheless, native speakers of all the languages have to be in charge of the validation process after automatic acquisition.
- Any automatic process of extraction can be replayed at any time with different parameters and adjustments in the method. This fact is very important because we will see that automatic extraction is largely dependent in terms of quality on the already existing and validated knowledge. Replaying extractions will improve the initial result only at the cost of cheap additional computer resources.

ICD10 as a Knowledge Source

The author's group has designed two complementary methods for automatic multilingual lexicons extraction from a corpus of texts. These methods have been experimented on the ICD10 classification which is currently available in numerous languages (more than 50 translations are underway [v]). ICD10 is an international classification of diseases primarily designed for statistical and epidemiological purposes[vi]. It acts as a common repository for international comparisons. Many countries have adopted ICD10 or some variant of it. Despite these variations, links between different national data bases are feasible using correspondence tables.

The major advantage and nearly unique in respect to other medical language sources is the fact that the ICD10 classification is translated in multiple languages. The author's group is currently working in English, French, German and Swedish, and this list will soon be extended. The translations are done manually by national experts, generally with care and search for quality. The present approach is a kind of reverse engineering which was certainly never anticipated by the authors of ICD10 and its translators : to extract the concepts which are used in the expressions together with their annotations in multiple languages.

Facts about ICD10

In order to give a better idea of the lexical content of ICD10, table 1 provides the most important figures as extracted from the source version. Those figures are subject to modifications because different interpretations of the recommended rules are possible. Nevertheless, they are representative of the reality.

Number of expressions (systematic): 12317
Number of words: 46701
Number of stop words: 32346
Number of different words (without multiple occurrences): 4777
Number of descriptors expressions: 9456
Number of additional different words: ~ 4000
Total number of words (without alphabetical part): ~ 9000
Estimated number of words, including alphabetical part: ~20000

Table 1: Figures about the lexical content of ICD10.

Extraction Strategies

Two complementary methods have been implemented for extraction of multilingual information. The first one is the simplest, but it clearly demonstrates the feasibility of automatic extraction. It relays on direct comparison of corresponding expressions in two different languages with an attempt to match words. The second one is based on distribution of words in the whole classification. It is less intuitive but it brings a high potential of good matches.

In both methods, a pre-processing phase is necessary with two specific goals : first, to eliminate the stop words (determiners, prepositions, conjunctions, etc.) ; second to transform any word to its basic form, independently of gender and number.

First Method : Direct Comparison of Expressions

The first method of extraction is based on the comparison of two expressions related to the same ICD10 code. Working on only two languages instead of all considered languages is not a limitation because subsequent runs will be possible with the different pairs of languages. Different runs are designed to successively extract linguistic knowledge.

The initial run is trivial : it takes all the expressions made of a single word in both languages and creates pairs of such words. Each pair is treated here as a new concept (labelled by only an internal number at this moment) with its annotation in two languages. When working with the English and French versions of ICD10, nearly 500 pairs have been found like *wheezing / sifflement*, *heartburn / pyrosis* or *headache /céphalée*. The ICD10 code is kept with each pair as the semantic justification for the final merge with multiple languages.

The second run deals with two-word expressions in both languages, from which one word is already paired from a previous run, from another method or from a manually written dictionary. Considering the pairs of second words when the first one is subtracted, numerous candidates emerge. This run may be played many times in sequence because the pairs discovered during one run open the way for new discoveries in the next run. Up to seven iterations have been productive of new pairs. Again, some hundreds pairs are found in this way, amongst them in English / French : *laryngeal / laryngé, tonsillar / amygdalien* or *mucocutaneous / cutanéo-muqueux*. It should be noted at this point that the method makes no hypothesis on the word category and consequently that a noun may be paired with an adjective. For example it has been found the following pair : *mumps / ourlien* where the English *mumps* is a noun and the French *ourlien* is an adjective. The ICD10 entry which has been used for that is the code B26.1 *Mumps meningitis / Méningite ourlienne* with the pair *meningitis / méningite* already known from a previous run.

The third run compares three-word expressions in one language with two-words expressions in the other one, asking that the third word is a stop word in middle position like a preposition used for noun complements (*of* in English or *de* in French). This is clearly a way to compare a noun complement expression in one language to a adjectival expression in the other language.

Further runs are designed to retrieve pairs of words. The process is more complex with larger expressions. The basic rule is to isolate words in expressions which are the last unknown words in both language. By chance, the mean number of words by expressions (after removal of stop words) being below 5, the capture of nearly all ICD10 words is feasible.

Second Method : Pattern of Co-occurences

The second method starts from any pair of expressions sharing the same code in two languages and looks for all the present words. Then, for each word, it builds the list of all other expressions of the classification having this same word : it is the pattern of co-occurrence of this word in the ICD10 classification.

Taking one word in an expression of the first language and all the words in the expression of same code in the other language gives a number of pairs. Each word of a pair has a pattern of co-occurrence to be compared with the pattern of the other word in its pair. A score of the number of exact match is determined as it may be seen on figure 1. In general, the score is high for one pair and

low for all the other pairs. The high score pair is considered as valid and its words are removed from their expression. The process continue with the rest of the expression. This strategy is generally efficient, but it needs some fine tuning. Nevertheless, the harvest is a substantial reflect of the reality and the results are improved from the first method. In addition, this method does not require multiple runs.

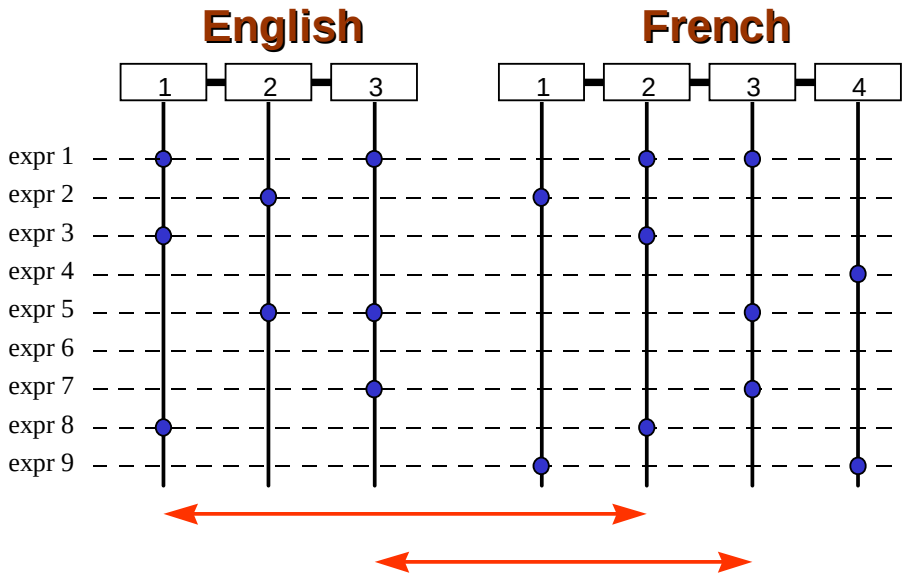

Figure 1: Extraction from the study of pattern of co-occurrences in two translations of ICD-10. The English word #1 matches with the French word #2. Words #3 are also in correspondence. The method looks for high match, but exact correspondence is not required.

This method will also take advantage of existing dictionaries which have been validated. In fact any known pair from an initial dictionary is taken as fixed and limits immediately and strongly the degree of freedom of this method and its associated combinatorial explosion. The larger is the initial dictionary, the more efficient is the retrieval rate of this method.

Results and Discussion

In order to further improve both methods, it has been decided to shift the work from words to the underlying concepts. This means mainly that when a morpho-semantic decomposition of words is available, the word is replaced by the underlying concepts before applying any one method. For that purpose we are using a method of word decomposition into their stems which has been already presented elsewhere[vii, viii]. For example, the word *gastroduodenoscopy* is transformed into the words *stomach*, *duodenum* and *endoscopy*, each word representing an underlying concept. This same process applies also on simple words like *gastric* which is also converted to *stomach*. This process of semantic aggregation relies on the existence of dictionary of word decomposition designed and implemented in the author's group with more then 8000 entries.

The present method has been applied on the French and English languages and provides a bilingual dictionary of nearly 10000 words. A validation of the results has been performed on a sample of 500 pairs of words and has shown preliminarily that the established correspondences are correct in more than 98% of the cases. Further validations are underway. The necessity of a human proof-reading is clearly established before reuse of such dictionaries in other applications. But the above high score of correct pairs facilitates the final validation process.

Conclusion

Semi-automatic linguistic knowledge acquisition is a real must because NLP developments are strongly limited by the non-availability of good and extensive lexicons. Moreover, it has been learned

that such lexicons are better build on top of conceptual models in order to insure the coherence of a multilingual approach, and to enforce their use by NLP tools. The present experiment demonstrates such an approach. It has been successful, because words have been segmented in parts which are direct pointers to concepts of the domain. The underlying semantic decomposition gives strength to this method, when purely lexical approaches have failed. This semi-automatic extraction scheme has certainly a potential for further development in order to acquire linguistic knowledge from textbooks.

References

[i] Lindberg DAB, Humphreys BL, McCray AT. *The Unified Medical Language System*. Meth Inform Med 1993, 32: 281-291.

[ii] *Multilingual Glossary of technical and popular medical terms in nine European Languages* http://allserv.rug.ac.be/~rvdstich/eugloss/welcome.html

[iii] Baud RH, Lovis C, Rassinoux A-M, Scherrer J-R. *Alternate Ways for Knowledge Collection, Indexing and Robust Language Retrieval*. To appear in: *Proceedings of the Fourth International Conference on Medical Concept Representation*, Jacksonville FL, 1997, pp 81-93.

[iv] Baud RH, Lovis C, Rassinoux AM, Michel PA, Alpay L, Wagner JC, Juge C, Scherrer JR. *Towards a Medical Linguistic Knowledge Base.* In: Greenes RA et al. (eds.), proceedings of MEDINFO 95. Alberta: HC&CC, 1995: 13-17.

[v] Private communication with a responsible officer at World Health Organisation WHO, Geneva, Switzerland

[vi] *International Classification of Diseases*. Different versions. World Health Organization, Geneva, Switzerland.

[vii] Lovis C, Michel PA, Baud RH, Scherrer JR. *Word Segmentation Processing : A Way to Exponentially Extend Medical Dictionaries*. Proceedings MEDINFO'95, R.A Greenes et al eds, 1995 IMIA, pp 28-32.

[viii] Lovis C, Baud RH, Rassinoux AM, Michel PA, Scherrer JR. *Building Medical Dictionaries for Patient Encoding Systems: A Methodology* In proceedings of Artificial Intelligence in Medicine Europe, Keravnou E & al. (eds.), Springer Verlag, 1997, pp 373-380.