A Light Knowledge Model for Linguistic Applications

Robert H Baud, PhD, Christian Lovis, MD, Patrick Ruch, MS, Anne-Marie Rassinoux, PhD.
Medical Informatics Division, University Hospital of Geneva, Switzerland

Content extraction from medical texts is achievable today by linguistic applications, in so far as sufficient domain knowledge is available. Such knowledge represents a model of the domain and is hard to collect with sufficient depth and good coverage, despite numerous attempts. To leverage this task is a priority in order to benefit from the awaited linguistic tools. The light model is designed with this goal in mind.

Syntactic and lexical information are generally available with large lexicons. A domain model should add the necessary semantic information. The authors have designed a light knowledge model for the collection of semantic information on the basis of the recognized syntactical and lexical attributes. It has been tailored for the acquisition of enough semantic information in order to retrieve terms of a controlled vocabulary from free texts, as for example, to retrieve Mesh terms from patient records.

Limits of present situation

The last three decades have seen large research developments in an attempt to build nomenclatures for the medical domain. Natural extension of these tasks is the elaboration of a comprehensive model of the medical domain. The major goals [1] are: indexing of documents, decision-making support, medical order control and monitoring, inference techniques, etc.

The main efforts are Snomed [2, 3, 4], Read terms [5, 6] and Galen model [7, 8] with similar order of magnitude regarding the coverage; this means a rather general model of medicine involving tens of thousands concepts or terms. Other approaches reach a comparable extension, though the foundations may be rather different. The intention of Cimino with a Medical Entities Dictionary [9, 10] was directed to the construction of a controlled vocabulary in the medical domain. The UMLS Mesh system [11], with a hierarchy of more than 33000 terms, is dedicated to literature indexing in Medline, but may be helpful in other contexts. The frame approach acting as an interlingua between different controlled vocabulary is also worthy to be mentioned [12].

The lesson of the nineties is that a complete model of the medical domain is hard to obtain with the presently mastered techniques of knowledge representation. Rector has cleverly presented this point of view in a recent paper [13], which is rather to be understood as a contribution to the track "lessons from the past and vision of the future." Nevertheless, recent information from private companies clearly shows that huge investments are underway, and that the advent of new modelling tools will change the scene in a near future.

Despite the association of Snomed and Read in pursuing this quest for the Holy Grail, with the explicit goal of designing and publishing a formal medical representation [14], with a positive perspective of success at the cost of substantial manpower resources, it becomes extremely important to start alternative approaches. One track to explore is the feasibility of a simpler model with general coverage of the domain, but less ambitious considering the power and depth of representation. Such a model shall be mainly guided by linguistic constraints rather than knowledge representation principles at the point of view of priority management. It will be legitimated by short-term results. In the authors' context, this means achieving a workable degree of document indexing with the Electronic Patient Record (EPR) and facilitating Medline access with translation of French queries, copied-pasted from the patient record. Relevant references exist in [15].

Another key characteristic of the modeling method is its multilingual capability. Numerous countries (especially in Europe, but US is directly concerned with Spanish) handle non-English medical records. Though basically the knowledge is language independent, this is not necessarily true for the acquisition methods and the availability of large size lexicons. Companion papers to the present one insist on this aspect [16, 17]. The importance of multilingual tools for knowledge acquisition is recognized [18].

A light knowledge model

The goal is the extraction of medical terms from natural language free texts, as found in the electronic patient record. Such terms are known from different sources having different names depending on their authors: an ICD expression when dealing with diagnosis; a Mesh term when indexing documents for retrieval; a Snomed entry when working with this

1067—5027/01/$5.00

nomenclature; an entry in a controlled vocabulary file from other pragmatic approaches.

The modelling approach here presented is a method aiming at the representation of the knowledge extracted from a sequence of terms as found in a terminology, in a formal manner compatible with conceptual graphs. The extraction process may be automatised, as soon as enough syntactical, lexical and conceptual information has been collected, as shown in the assessment of the method.

To illustrate a quite common situation when querying a corpus of text, one can look at the multiple entries of a UMLS CUI (figure 1). This is a typical situation to be solved.

Aortic valve insufficiency
Insufficiency, aortic valve
Aortic insufficiency
Aortic incompetence
Incompetence, aortic valve
Incompetency, aortic
Incompetent, aortic valve
Regurgitation, aortic valve
AR – aortic regurgitation

Figure 1: C0003504 Aortic Valve Insufficiency. There are 37 proposed variants for this term in English. A few typical terms are presented here.

The model of representation of terms is a knowledge scheme with the following features [16]:

- Ability to represent all terms; robustness regarding the unknown words.
- Ability to cope automatically with new terms or variant free text expressions. This implies the possibility to expand the model using automatic acquisition techniques, without mandatory heuristic interventions by human.
- Recognition of synonym terms as far as they differ only by morphological, syntactical variants or etymological roots. Synonyms should finally point to the same representation as the source or preferred term. Example: *spondylodiny, chronic* and *permanent vertebral pain.*

Terms representation

In order to achieve terms representation there are multiple steps to follow. The authors have implemented the methodology presented here, ready to be used for the management of the EPR. This section enumerates the main steps with limited implementation details. For more information, see the bibliographical references.

Step 1: Tokenisation of the input term.

This step seems simple to the novice, but it is just the appearance; it may be complex for different languages, because natural languages are really not formal languages and numerous exceptions and unexpected situations are the rule. It should be well mastered, because this step brings information from the lexicon into the model, and what is missed here is missed for ever. The problems to solve are :

- Recognition of morphological variants of words which are present under basic form in the lexicon: plural forms in all languages, gender form in French, German or Italian, case forms in German or Finnish, etc.
- Recognition of significant stems [19] or parts of words, quite frequent in the medical domain.
- Spell checking and orthographic correction [20, 21], which may account for a few percent of unrecognized words when not present, depending on the quality of medical texts.

Current results on this technique are excellent for morphology variants discovery, stem decomposition and spell checking. However, when coping with clinical texts, absence of word in the dictionary, specific abbreviations and units left us with 10% of unknown words. Half of them have been shown to disappear using a local dictionary. The rest necessitates an drastic improvement of the coverage of the domain by the dictionary.

Step 2: Disambiguation.

Ambiguities are often underestimated by group of people working with a prototype lexicon (typically 5000 entries). The reason is evident: ambiguities augment with lexicon size and this increase is not linear for sure. Therefore, this step is a must. Tagging techniques are known and their usage in the medical domain exist elsewhere [22].

- Resolution of syntactic ambiguities, when the same word has more than one syntactic category.
- Resolution of ambiguities at a word level, when the same word may have two different conceptual attachments depending on the context.

As shown in the above mentioned reference, our rate of success at this level is approximately 99%, this means less than 1% of ambiguities.

Step 3: Parsing the term.

Parsing is the structural step in the middle of the process. Parsing transforms a sequence of words in noun phrases in a graph-like structure, based on syntactic arguments. The authors have developed a shallow rule-based parser [23] for this task. It has the advantage of being robust and is always in position to extract short noun phrases from complex sentences:

- Parsing the medical sentences.

- Extracting noun phrases and building their syntactic structure.

Step 4: Semantic tagging.
Independently of step 2, often based on a set of syntactic tags, this step is clearly a semantic approach. We have explored a tag set [24] closely related to the UMLS Semantic Net concepts. Results are good and the tagging of our entire lexicon is underway. We propose here to exploit this tag set with a distinct purpose: to define semantic attachments for rule-based modeling of terms.

The principle is the following: having obtained structural links under the form of syntactic relationships between words in a sentence, we want to transform them into semantic links. To do that, we need high-level tags in order to categorize the attachment of the syntactic links. The set of rules will be principally based on these categories.

- Adding semantic tags for each word entry in the input sentence, based on the Semantic Net

Step 5: Syntax-driven modeling.
This step is the key to the light model presented here: to transform the remaining syntactic links between words into semantically relevant relationships. In this way the initial sequence of words, structured by the parser, is transformed in a graph-like structure, representing the input term.

For example, in the presence of *cerebral haemorrhage* giving after syntactical parsing and tagging (*papr* is a tag meaning *pathological process* and *loc* is a tag meaning *body location)* :

[cl_Haemorrhage | papr]-(HasAdj)-[cl_Brain | loc]

With a general and frequently used rule like *"if a papr is linked to a loc by the syntactical link HasAdj, then the semantical link is HasLocation"*, the new form is:

[cl_Haemorrhage]-(HasLocation)-[cl_Brain]

- Transform with a set of rules syntactical links into semantically relevant relationships.
- Implementing an engine for application of the rule set and production of a language independent knowledge representation of the initial term.

Assessment of the method

In order to measure the feasability of this method, we experimented with an automatic knowledge acquisition process from an existing corpus of terms. Working on 12'316 systematic terms of ICD10, we found 93'567 occurences of words. After removing the stop words, except prepositions, we are left with 61'404 words. Of these words, 28'642 occurences were semantically tagged from our current dictionary.

As a demonstrative example, we focused on pathological process (papr), which is explicitely tagged in order to acquire information on their locations (loc). We found 684 expressions with both tags. From them, 240 expressions have a location specified by an adjective and 406 by a noun.

A manual check of the above 240 pairs found (noun/papr followed by adjective/loc) shows that 75% of them were correct. Incorrect ones are due to a long distance between the matching words. Furthermore, limiting the distance of words in the source term to 3 raises the score to 100% for the remaining 112 terms, amongst which are: pleural effusion (J90), subdural abscess (G06.2), cortical necrosis (N17.1), anal prolapse (K62.2), stricture of ureter (N13.5), etc.

This absence of noise, together with a substantial number of retrieved terms, is considered as an ideal situation for automatic knowledge acquisition. However, missing pairs (silence) are numerous and essentially due to untagged words, a problem easily solved by additional manpower resources working on basic lexicon, and incorrectly tagged words.

Coping with similar terms

It is well known that there are numerous wordings for the same term. This is true whatever the source of terms or the language. It can be said that too often the terms are hidden by the words in practical medical reports. This will remain true as long as NLP tools are not substantially more powerful. In the example of figure 1, one can imagine more prosaic formulations like "*aortic valve with chronic insufficiency*" or "*regurgitation at the level of the aortic valve*".

To think that physicians could be educated for usage of a standard vocabulary is not a good idea. There is no hope in this direction. Text are often written in a hurry, and voice recognition systems will shift the main actors from written to oral texts, which are known to contain more irregularities of the language. This fact will worsten the problem in the future.

One solution to this problem lies in the capability of the term analysis technique to render equivalent final representation of synonym expressions, whatever are the discrepancies at the surface language level. This can be achieved in two complementary phases: at a word level playing with syntax and morphology until matching to a source term is found; at a semantic level on the basis of explicit links between near concepts. The first phase is the main one and it is expected to resolve 80% or more situations. It is illustrated now with the example of the *spondylodiny*.

First expression: spondylodiny, chronic
The first step retrieves 3 words: the prefix *spondyl*, the suffix *odiny* and the adjective *chronic*. The words are linked to the concept they represent:

spondyl -> cl_Vertebra,
odiny -> cl_Pain
chronic -> cl_Chronicity.

Steps 2 is skipped in this case. Step 3 uses a shallow parsing method to recognize the two syntactical links:

[cl_Pain:Odiny]-(HasPrefix)-[cl_Vertebra:Spondyl]
-(HasAdj)-[cl_Chronic:chronic]

Step 4 gives the tag *loc* (meaning *Body location*) to *spondyl*, the tag ss (meaning *Sign and symptom*) to *odiny* and the tag temp (meaning *temporal attribute*) to *chronic*. The result is:

[ss]-(HasPrefix)-[loc] and [ss]-(HasAdj)-[temp]

Step 5 replaces the concept by their tags and looks for a matching syntax driven rule from a predefined set of rules. A typical rule has the form:

if [ss]-(HasPrefix)-[loc] then [ss]-(HasLocation)-[loc]

The effect of the rule is the replacement of the syntactic link by a semantic one, giving the final result:

[cl_Pain]-(HasLocation)-[cl_Vertebra]
-(HasTemporal)-[cl_Chronic]

Finally, one is able to store and retrieve the term representation indexed by any of its constituent concepts, in this case under *Body location* and under *Sign and symptom*.

Second expression: permanent vertebral pain

The same process is applicable. The first step of tokenisation retrieves 3 words: the adjective *permanent*, the adjective *vertebral* and the noun *pain*. The words are linked to the concept they represent:

vertebral -> cl_Vertebra,
pain -> cl_Pain
permanent -> cl_Permanence.

Step 3 uses a shallow parsing method able to recognize the two syntactical links:

[cl_Pain:pain]-(HaAdj)-[cl_Vertebra:vertebral]
-(HasAdj)-[cl_Permanence:permanent]

Step 4 retrieves the tag *loc* (meaning *Body location)* to *vertebral*, the tag *ss* (meaning *Sign and symptom)* to *pain* and the tag *temp* (meaning *temporal attribute*) to *permanent*. The result is:

[ss]-(HasAdj)-[loc] and [ss]-(HasAdj)-[temp]

Step 5 replaces the concept by their tags and looks for a matching syntax driven rule from a predefined set of rules. Not the same rules as above in the first example expression will be necessarily used. An identical result is obtained as above:

[cl_Pain]-(HasLocation)-[cl_Vertebra]
-(HasTemporal)-[cl_Chronic]

Power and limitation of the method

The above demonstration looks simple, but the reality is more complex. We know about a few limitations making life difficult and we want to discuss them.

This light model approach is based on the existence of a lexicon with syntactic attributes (for the tokeniser, the tagger and the parser) and conceptual attachment (for word concepts to solve synonyms and for high level tags before applying the rules). Such a lexicon should have a broad coverage of the language with a minimum of 20000 basic form entries (not including morphological variants). This is the size of the authors' lexicon on which they develop the present experiment.

This approach assembles multiple NLP techniques for which one needs a good and efficient implementation, not easy to master. Amongst the basic tools involved in the process are: morphologic word analysis, morphosemantem decomposition, disambiguation, parsing, semantic tagging and rule based resolution of links. Nevertheless, such tools are well known by the specialists for decades now.

The main question is then: what is the rate of success of the method or what kind of failures is expected? Wrong positive matches seem not to be a major problem. The medical language is a scientific language able of acute precision when needed (like operating room reports). In case of wrong match improvement is expected by augmenting the set of matching rules. Another solution is the refinement of the existing rules, either by more detailed high-level tags, or by specifying specific occurrences of words (a prepositional attachment is different for *with* or *without*).

On the other end the missing matches may be a problem. A miss occurs when a term has not been recognized in a given expression, despite it is present, probably hidden behind a too complex sentence, unrecognized by the computer process. Our current experience is clearly positive as long as we work with rather simple noun phrases and limited to 5 or 6 words. The potential for amelioration is high and this limit should disappear in the future.

A big advantage of a rule-based approach is the fact that the complexity of growing the model becomes somewhat linear: doubling the set of rules would double the efficiency of the system. This may not be true in the long term, but there are long expected developments in this direction before reaching other limits.

Future developments and conclusion

Different axes of future developments are easily sketched now: augmenting the coverage of the lexicon in at least two languages: English and French, possibly others depending on co-operations; refining the conceptual information in lexicon; finalizing the high level semantic tags and stabilizing the set of tags; working on the parsing of more complex noun

phrases and possibly verb phrases; refining the modeling rules for differentiation of similar terms when necessary.

Knowledge extraction in the form of regular co-occurences is feasible thanks to the light model. We have shown how to handle lists of terms in a classification for extracting knowledge. This process is dependent on a language lexicon and expert tools applying NLP techniques.

However, it has also been shown that low noise is a realistic target when performing knowledge extraction. This is a necessary condition for automatic knowledge acquisition. This approach is expected to provide a significant advantage compared to manual or heuristic approaches. The name "light model" is now explained: simply tagging words in a lexicon leads to sound knowledge extraction. Finally, needless to say, more sophisticated models are certainly compatible with lighter models.

REFERENCES

[1] Sittig DF. Grand challenges in medical informatics. J Am Med Inform Assoc 1994; 1: 412-3.

[2] Côté RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L, eds. The systematized nomenclature of medicine: Snomed International. Northfield, Illinois: College of American Pathologists, 1993.

[3] Web site at: www.snomed.org

[4] Rothwell DJ. Snomed-based knowledge representation. Meth Inform Med. 1995; 34: 209-213.

[5] O'Neil MJ, Payne C, Read JD. Read Codes version 3: a user led terminology. Meth Inform Med. 1995; 34:187-92.

[6] Schulz EB, Price C, Brown PJB. Symbolic anatomic knowledge representation in the Read codes version 3: structure and application. J Am Med Inform Assoc. 1997; 4:38-48.

[7] Nowlan WA, Rector AL, Rush TW, Solomon WD. From terminology to terminology services. J Am Med Inform Assoc 1994; (Symp supplement): 150-4.

[8] Rogers J, Rector A. The GALEN ontology. In Medical Informatics Europe (MIE96). Copenhagen: IOS Press: 1996; 174-8.

[9] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of large controlled medical terminology. J Am Med Inform Assoc 1994; 1: 35-50.

[10] Cimino J. Desiderata for controlled medical vocabularies in the twenty-first century. Meth Inform Med 1998; 37 (4-5): 394-403.

[11] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Meth Inform Med. 1993; 32(4):281-291.

[12] Masarie FE, Miller RA, Bouhaddou O, Nunzia BG, Warner HR. An interlingua for electronic interchange of medical information: using frame to map between clinical vocabularies. Procs 17th Annual Symposium on Computer Applications in Medical Care. Washington, DC: McGraw-Hill, 1993: 829-833.

[13] Rector AL. Clinical terminology: Why is it so hard? Meth Inform Med 1999; 38: p239-252.

[14] Spackman KA, Campbell KE, Côté RA. Snomed-RT: A reference terminology for health care. J Am Med Inform Assoc 1997; (Symp. Suppl): 640-4.

[15] Cooper GF, Miller RA. An experiment comparing lexical and statistical methods for extracting Mesh terms from clinical free text. J Am Med Inform Assoc 1998; 5: 62-75.

[16] Baud RH, Lovis C, Ruch P, Rassinoux A-M. An initiative to develop an international recipient for a multilingual dictionary in the medical domain. Submitted to IJMI, 2001

[17] Baud RH, Lovis C, Ruch P, Rassinoux A-M. Conceptual search in electronic patient record. Medinfo 2001, London UK.

[18] Ceusters W, Buekens F, De Moor G, Waagmeester A. The distinction between linguistic and conceptual semantics in medical terminology and its implication for NLP-based knowledge acquisition. Meth Inform Med 1998; 37 (4-5): 327.33.

[19] Lovis C, Baud RH, Michel P-A, Scherrer J-R. Morphosemantic decomposition and semantic representation to allow fast and efficient natural language recognition. J Am Med Inform Assoc 1997; (Symposium Supplement): 873.

[20] Ruch P, Gaudinat A. Comparing corpora and lexical ambiguity. Proc of the ACL Workshop on comparing corpora. Hong-Kong 2000. ACL Eds.

[21] Ruch P, Baud RH, Geissbuhler A, Lovis C, Rassinoux A-M, Rivière A. Looking back or looking around: comparing two spell checking strategies for documents edition in an electronic patient record system. Submitted to AMIA Fall Conference 2001.

[22] Ruch P, Wagner JC, Bouillon P, Baud RH, Rassinoux A-M, Scherrer J-R. MEDTAG : Tag-like semantics for medical document indexing. J Am Med Inform Assoc 1999; (Symposium Suppl): 137-41.

[23] Baud RH, Rassinoux A-M, Ruch P, Lovis C, Scherrer J-R. The power and limits of a rule-based morpho-semantic parser. J Am Med Inform Assoc 1999; (Symposium Suppl): 22-26.

[24] Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Medical document Anonymisation with a semantic lexicon. J Am Med Inform Assoc 2000; (Symposium Suppl): 729-33.