Morpho-Semantic Parsing of Medical Expressions

Robert H Baud, PhD, Christian Lovis, MD, Anne-Marie Rassinoux, PhD, Jean-Raoul Scherrer, MD
Medical Informatics Division, University Hospital of Geneva, Switzerland

The task of editing, indexing, storing, and retrieving medical expressions within medical records remains the main objective for the years to come. Therefore, the need for a parser with semantic capabilities able to robustly extract an essential part of the knowledge embedded in the medical record is paramount. The minimal requirements before considering clinical trials are that such a system has to be in position to handle any source of medical information and to conveniently grasp the main key concepts with low silence, good recognition of modalities and acceptable noise. This paper shows that the potential of morpho-semantic parsing is high to meet these conditions. This technique is an important complement to the traditional lexical approach and to expression-oriented systems like controlled vocabularies.

INTRODUCTION

Natural Language Processing (NLP) tools are foreseen as the vector of the next revolution in medical informatics, just as mouse and graphical tools have changed the world in the nineties. The fact that software manufacturers are presently investing hundreds of man years on this topic is a clear sign. The medical domain will be one of the first targets for new products like voice control and text understanding. This is clearly a paradigm shift for health systems.

Electronic Patient Records are currently under development at multiple sites all over the world and quite a number of such systems are routinely used. Before such a huge quantity of medical texts, the question is raised of what NLP functionalities are required to properly grasp the content and knowledge embedded in texts, what kind of indexes and key concepts are to be extracted, what degree of detail is convenient for a formal representation, and more generally what kind of semantic processing is foreseeable in the couple of years to come. It appears clearly that the simple storage of medical texts "as a sequence of characters" is undoubtedly useful. Nevertheless, it misses the basic feature of organising information. The ever increasing quantity of information collected for each patient makes this problem even more important than it has ever been. Medical texts and the electronic patient records constitute mostly an unstructured corpus of information. Despite that it may be retrieved in chronological order and by patient identifier, it is generally not reachable on the basis of specific concepts hidden in the mass of characters and words. This is the main problem to be solved today and NLP tools are sought as a possible solution.

THE VALUE OF MORPHO-SEMANTEMS

The authors' group has been long active with different NLP techniques in the medical domain. A key idea emerged a couple of years ago about the advantage and usefulness of morpho-semantic expressions. Many papers were published on the subject in the seventies. But the best ideas are not really of value until they are efficiently implemented in the appropriate context. This is, may be, what is new about morpho-semantic expressions today and this paper is a description of a full range morpho-semantic parser of natural language expressions in the medical domain.

English or French are not really considered as compositional languages, in the sense that they do not really use compound words made from two or more other words as issued from a basic lexicon. On the contrary, German and many other languages are clearly compositional and they currently build compound words. Figure 1 is a simple comparison between these three languages.

Lexicons are usually built on the following grounds: the basic entry or unit of work is the simple word usually separated from another word by a blank character. On the contrary, the present approach defines the morpho-semantic as its basic unit of work and not the word. Typically, it decomposes a number of words into their meaningful units (e.g. prefixes, suffixes, etc.). This means that entries in the lexicon correspond to semantic atoms. In general, a morpho-semantic points to an elementary concept and it is not linked to a combination of two or more concepts of the domain. This is quite a fundamental shift from the point of view of language representation. We clearly consider that such an
approach is complementary from the point of view of other active actors in this field in their expression of the requirements for clinical vocabulary.

<table>
<thead>
<tr>
<th>English: Aortic valve insufficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>German: Aortenklappeninsuffizienz</td>
</tr>
<tr>
<td>French: Insuffisance de la valvule aortique</td>
</tr>
</tbody>
</table>

**Figure 1:** Compared compositionality between English, German and French.

In order to make a clear distinction between different approaches when building a lexicon, three degrees of granularity are possible:
- expression lexicons (professional dictionary)
- single word lexicon (home dictionary)
- morpho-semantem lexicon (computer dictionary)

A given lexicon is ordinarily limited with respect to the above possible entries. Single word entry dictionaries are the most common for home dictionaries. Entries are classified in alphabetical order and accessed by a single keyword. Under the entry, quite a lot of information may be given, in particular typical expressions linked to this entry.

<table>
<thead>
<tr>
<th>Traditional lexicon:</th>
</tr>
</thead>
<tbody>
<tr>
<td>otitis, noun -&gt; composite concept</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Morpho-semantem lexicon:</th>
</tr>
</thead>
<tbody>
<tr>
<td>oto, prefix -&gt; cl_Ear</td>
</tr>
<tr>
<td>itis, suffix -&gt; cl_Inflammation</td>
</tr>
</tbody>
</table>

**Figure 2:** Transition from a usual lexicon to a morpho-semantem lexicon: an entry of the lexicon is now an atom to the semantic point of view

Professional dictionaries are built with expressions. The Metathesaurus from UMLS is a repository of multiple words expressions. Each entry is referring to some conceptual entities which are usually complex. The same is true for lists of findings. Depending on the underlying semantic model of representation, such an entry is a composite concept pointing to two or more elementary concepts.

Morpho-semantem lexicons are not common. In fact, they are not good for reading by human beings, but they are essentially relevant for computer applications. As shown in Figure 2, their entries are generally referring to an elementary concept and therefore, they are particularly useful for automatic extraction of knowledge from medical texts, which is characterised by the frequent use of compound word forms. This is a main difference in order to demonstrate the potential of this method.

**A MORPHO-SEMANTIC PARSER**

Multiple parsers have been developed since the sixties and multiple authors have published excellent reviews on this topic. Morpho-semantic parsing is a variant of high interest in the medical domain. Capabilities of a morpho-semantem parser of natural language expressions are presented in Figure 3. They are the basic infrastructure for useful tools in a clinical setting. Not a single step in the above list may be omitted. In the next sections, we want to examine each step in turn.

- to recognise all the words in their basic form,
- to decompose all the compound words into their morpho-semantems,
- to link words to the concepts of an underlying model,
- to build an unambiguous parse tree of the expressions.

**Figure 3:** The basic steps for practical handling of medical expressions in a clinical setting.

**Recognising basic forms**

Recognition of words in their basic form has been realised through a rule based system. There are 3 sets of rules to be rewritten for each language and this has been done for English, French and German. We ended with more or less 100 rules in each case. The sets are the following: 1) rules for singular from plural and declined forms; 2) rules for guessing adjectives; 3) rules for guessing nouns. Presently, verbal forms are not handled and we concentrate on noun phrases. The guessing rules are based on typical endings and possibly help to deal with unknown words. They allow in theory to shorten the size of the lexicon. In addition, a lexicon of irregular forms exists.

**Morpho-semantem decomposition**

The basic idea of morpho-semantem decomposition is the fact that numerous words in the medical domain are composed from two or more concepts. Is it not more natural to work at the level of the natural single concepts when analysing sentences, or working at the level of words? What is the most basic unit? The authors' group has soon privileged the semantical approach and therefore has elected the morpho-semantem decomposition as a favourite.
scleritis -> Sclera + Inflammation
tleojejunostomie
   -> Ileum + Jejunum + Stoma
musculoskeletal -> Muscle + Skelette

Figure 4: 3 typical morpho-semantic decompositions. Decomposition into morpho-semantic terms has been described elsewhere. Figure 4 gives a couple of typical decompositions which have now been automated to a full proof algorithm.

Building parse tree
There are strong incentives for a parsing process which is exceptionally robust when confronted with medical texts. The main two reasons are: medical texts are often written in a hurry and they have plenty of abbreviations, jargon and misspelling errors; the lexicon coverage is a never-ending task and unknown words are foreseeable at any time. Any analysing system should be able to cope with multiple situations of unrecognised entries with minimal loss to the rest of the text. Typically in the presence of an unknown word in the middle of a sentence, the parser should correctly handle both remaining parts. Our design can handle any number of missing words and cope with the rest of the text.

After having analysed all the words of the input sentence in a sequence and extracted the available information from the lexicon, we have a linear list of words. Then we build a “linear graph” of all the words (see Figure 5). With this raw input, we trigger a rule-based parser with the following characteristics:

1. The parser is implemented as a set of rules designed to match predefined sequences of word categories in a sentence, like an adjective-noun rule or noun-conjunction(and)-noun rule,
2. Rules are ordered according to the specificity of the language, and are selected in that order,
3. Each rule examines in turn all sequences of two or three entries of the graph and may or may not be triggered,
4. Each rule specifies what kind of lexical checks have to be performed in gender and number on the selected entries, and fails if unsuccessful,
5. Each rule is possibly validated against a semantic model allowing, or not, its application in the present context; in the absence of a successful validation, the rule fails; in the absence of a semantic model, certain ambiguities will not be resolved and this step is skipped,
6. Each triggered rule simplifies the 2 or 3 selected entries of the graph into a single entry and the rest as a dependant sub-graph, thus decreasing the initial number of entries,

7. After the processing of all the rules (currently 40 rules), the initial graph has been transformed into a single entry graph giving evidence of full successful parsing,

8. Eventually, if more than one entry is left, it means the parsing has not been totally successful, then each entry may be considered as a partial successful graph.

This approach has the advantage of producing a parse tree in all situations. Of course, if no rule has been successfully triggered, the result for a sentence of N words is N parse trees of no use. But, usually when just one or two problems are found in a given sentence, we are left with two or three parse trees which are relevant as vectors of information. Figure 5 shows the situation of a typical expression from a medical note.

Linear graph:

(rel1) -> [no]
(rel2) -> [ulcerative]
(rel3) -> [recto]
(rel4) -> [sigmoid]
(rel5) -> [itis]

Parse tree:

[cl_InflammationProcess]
   -> (hasPrefix) -> [cl_SigmoidColon]
   -> (hasPrefix) -> [cl_Rectum]
   -> (hasAdj) -> [cl_Ulcerative]
   -> (hasAdverb) -> [cl_Absence]

Figure 5: Initial “linear graph” and resulting parse tree for the sentence “No ulcerative rectosigmoiditis.”

Linking to an underlying conceptual model
All the entries in our lexicon are pointing to a concept acting as a recipient for synonyms in a given language, and as a bridge of corresponding words between different languages. Automatic extraction of these concepts has been described in a previous paper by the authors. Figure 6 shows the usefulness of this approach in the medical domain.

EXTRACTING CO-OCCURRENCES
On behalf of existing parse trees of medical expressions stored under the form of conceptual graphs, the road is open for extraction of co-occurrences of pairs of concepts within a given corpus of text. We have selected for that ICD-10 because it is currently available in multiple languages and because it has a broad coverage with more than 70'000 expressions. From the language
point of view, the weakness is certainly related to the coverage of clinical situations, patient complaints, signs and symptoms, because this classification has been mainly developed for epidemiological purposes and it has been designed for disease descriptions and diagnoses.

<table>
<thead>
<tr>
<th>Concept cl_Eyelid:</th>
</tr>
</thead>
<tbody>
<tr>
<td>eyelid, singular noun</td>
</tr>
<tr>
<td>palpebral, adjective</td>
</tr>
<tr>
<td>blepharo, prefix</td>
</tr>
<tr>
<td>paupière, feminin singular noun</td>
</tr>
<tr>
<td>palpébral, masculin sing. adjective</td>
</tr>
<tr>
<td>blépharo, prefix</td>
</tr>
</tbody>
</table>

Figure 6: Assignment of a concept to synonym words in the lexicon for English and French.

The study of co-occurrences is an essential step when studying sources of textual knowledge. Amongst the four categories of knowledge to be acquired for a complete system, co-occurrences is the third one after lexicon acquisition and typology of concepts, the fourth one being the script based approach, also named conceptual schemata. The statistical approach to co-occurrences leads to semantic rules able to guide different NLP tools either in analysis of texts or in the generation process. Indeed, from the observed frequency of co-occurrences in a large corpus, it is possible to largely validate any new texts and to solve ambiguities about lexical constructions. Co-occurrences lead to co-locations or semantic rules or clusters, also called sensible statements in another model. Those semantic rules necessarily are correlated to the relationships between concepts in any model of the domain.

The algorithm for co-occurrence extraction is rather straightforward. First, each expression from a classification or each sentence from a report or a discharge letter is initially converted to one or more conceptual graphs using the morpho-semantic parser. Second, each graph is processed with examination of all pairs of concepts where one is dependent on the other. Finally, a filter is applied on the lexical relationship linking the two concepts. The main links which are selected are: noun to noun, noun to adjective, noun to prepositional group, etc.

Handling coordinations in medical expressions

When dealing with medical classifications, one is always surprised by the extensive usage of multiple conjunctions. This is due to the need to have a large coverage of the domain and because this goal is often reached by simple enumeration. The problem with conjunctions like and, or and punctuation marks like comma is the fact that they bring numerous ambiguities of scope, well-known to all linguists. It is the role of the parser to solve those ambiguities, as far as possible, knowing that a total disambiguation is an impossible task. In multiple situations (see Figure 7), when considering usual classifications like ICD, only a deep knowledge of the specific discipline leads to the final version.

| Allergic and dietary gastroenteritis and colitis. |
| Otalgia and effusion of ear. |
| Congenital malformation, deformations and chromosomal abnormalities. |

Figure 7: Typical ambiguities in scope due to conjunctions, more or less easily solved by a human reader.

Negations and modalities

When writing systematically the history of a patient, reviewing each system, the physician can write numerous sentences in the negative form. Indeed, the negative findings, explicitly stated, are of major importance here. This same remark may be extended to all kind of modalities expressed in the medical texts, like suspicion, uncertainty, hypothesis, etc. Failure to grasp any negation or modality may perfectly and easily lead to wrong conclusions.

Our parser has been strongly influenced in this respect. The rule-based approach privileges the recognition of negation expressed by adverbs (see Figure 5) and prepositions (like without). The design of concept is especially sensitive to any expression of the modalities and words are oriented to modal concepts whenever necessary (cl_Doubt, cl_Suspicion, cl_Hypothesis, cl_Absence). Care is also given to prefixes like a-, dys- or mal-.

ADVANTAGES AND LIMITATIONS

Morpho-semantic parsing brings several benefits compared to other approaches, but it shows also a few limitations. The authors' experiences emphasise the following advantages: 1) Morpho-semantic dictionaries allow endless acceptance of new composite words, despite their shorter size (60%) than other dictionaries; 2) Correct decomposition of compound words is achieved in more than 99% of the cases for English and French; 3) Finer analysis of sentence meaning; 4) Additional knowledge is harvested through the study of co-occurrences.

In addition, the following limitations have to be taken into account: 1) Morpho-semantic decomposition may be ambiguous; 2) Morpho-semantic terms may have context-dependent meaning.
difficult to disambiguate, due to de facto language evolution; 3) Rules for parsing may be too complex to group together in an appropriate order.

CONCLUSION
This paper shows the benefit of a morpho-semantic parser compared to traditional word approaches. It brings a simpler analysis and a better recognition of the underlying concepts which are the expression of the text. The implemented system has shown itself to be robust and to always bring partial results when full analysis is not possible. Being paired to a bottom up semantic model of concepts, it proves that it can provide considerably finer analysis of the corpus of texts, which is particularly useful in the medical domain.

References

764