The Power and Limits of a Rule-based Morpho-Semantic Parser

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

The venue of Electronic Patient Record (EPR) implies an increasing amount of medical texts readily available for processing, as soon as convenient tools are made available. The chief application is text analysis, from which one can drive other disciplines like indexing for retrieval, knowledge representation, translation and inferencing for medical intelligent systems. Prerequisites for a convenient analyzer of medical texts are: building the lexicon, developing semantic representation of the domain, having a large corpus of texts available for statistical analysis, and finally mastering robust and powerful parsing techniques in order to satisfy the constraints of the medical domain. This article aims at presenting an easy-to-use parser ready to be adapted in different settings. It describes its power together with its practical limitations as experienced by the authors.

THE MORPHO-SEMANTEM APPROACH

In a recent paper, the specificity of the morpho-semantem approach was highlighted [1]. The pioneers on this track have published their work already in the seventies [2, 3]. Another author was cited in this paper as initiator of this approach [4]. The basic idea is that if the goal is knowledge representation of the text, and this representation is based on concepts of the domain, the morphemes are better suited to match the concepts than the words. Many languages used in the medical domain are compositional, and therefore words are not the best unit of decomposition. An *ileojejunostomy* is an entry which aggregates 3 different concepts: the body part *Ileum*, the body part *Jejunum* and the surgical deed *Stoma*. Morpho-semantem decomposition is a technique able to decompose the words into their meaningful parts.

To illustrate this approach, let us discuss the variability of formulation allowed in most languages. What about an alternative formulation like *jejunoileostomy*? This is neither anatomically nor functionally the natural way to describe this surgical deed, but it is semantically understandable. Another text could be "*creation of an ileojejunal stoma*" which is perfectly correct, or "*stoma structure from ileum to jejunum*" which is acceptable. This versatility favors the solution with the highest granularity, which is best suited for re-conciliation of all entries.

STEPS FOR WORD RECOGNITION

Before parsing a text, different steps have to be taken in order to recognize words in a sentence. Though implementation dependent, especially from the lexicon design, these steps are basically the following:

- **Conversion of words to their basic form**. This step aims at recognizing any variant forms of a word (plural, feminine, case, adverb, verbs largely dependent on the language) and converting it to its basic form as contained in the lexicon; examples are: *heads-head, allergies-allergy, feet-foot, slowly-slow* in English, *nerveuse-nerveux* in French, *ikterischer-ikterisch* in German; the difficulty lies in the fact that sometimes more than a single basic form exists for a given word: in French *fils* is either the word for *son* or the plural form of *fil* meaning *thread.*
- **Retrieval of basic forms in a lexicon**. This step has to be particularly efficient for lexicons with more than 20000 basic form entries; a given basic form may appear more than once in the lexicon when having different meanings like *patient* as a noun or an adjective. When the word is retrieved, related information is made available: syntax, usage, semantic, etc.
- **Morpho-semantem decomposition**. This step deals with the separation of compound words into their components, which in turn have to be retrieved as basic forms in the lexicon. The elected strategy is to first retrieve the longest match working from left to right, until the word is successfully exhausted, satisfying all the constraints. In the case of failure, a shorter match is attempted on backtrack. The main constraint is on the word category like prefix at the beginning and noun, adjective or suffix at the end. Special care should given to intermediate characters between morphemes: a *dash*, stems like *e, es, er, en* in German, etc. At this level, processing is purely lexical.

1091—8280/99/$5.00

- **Term analysis**. This step is necessary to recognize multiple word expressions. In this context, a term is defined as an expression without punctuation or special characters. Two tasks are performed: first, retrieval of a multi-word lexicon entry, this means an entry with blanks; second, separation of the words in the expression. The former task is easier to achieve when only considering invariant multi-word expressions. The latter task is complicated by the fact that a non blank character may act as a word separator like an *apostrophe* in French. The problem of lexical changes within multiple words expressions, like plural forms or optional additional words, has to be solved at this level.
- **Sentence analysis**. This step introduces the handling of punctuation and special characters. Some punctuation marks mean the end of a sentence, but the comma acts usually as a separator within the sentence. The presence of any kind of parentheses should be taken into account, and coherent matching checked. The result is the decomposition of the sentence in a set of terms to be separately processed as presented in the above point.

TAGGING FOR DISAMBIGUATION

From the initial steps, we are left with an ordered sequence of words in their basic form with all their attributes. Figure 1 shows an example for the term "*frequent perforated otitis with purulent discharge*" in the form of a table. Semantic concepts are issued from the GALEN model through a model-based semantic lexicon as described by the authors [5].

The order of the entries is language dependent and not necessarily semantically relevant. The focus of such a term is the suffix *–itis* meaning *inflammation lesion*. It is in the fourth position in English. The focus is semantic information conveyed by the term and extracted through the analysis. In a slightly different term like *"purulent discharge due to a perforated otitis"*, the focus is on *discharge*. This point illustrats the importance of parsing, the main task of which is to re-order the syntactic distribution of words.

In reality, the situation is much more complicated than the one presented here as a single list of words, due to the fact that any word may have multiple meanings depending on the context. This may lead to a combinatorial explosion. If a term presents two words with three meanings and two words with two meanings and the rest with one meaning, which is not uncommon, we are left with 3 times 3 times 2 times 2 = 36 combinations. Most probably, only one combination is meaningful! A combinatorial explosion is not at all welcome in the analyzing process because it leads to unpredictable and unacceptable process time. In order to solve this problem, we advocated tagging techniques [6], which are typical approaches in this direction.

morpheme	syntax	semantic
frequent	adjective	cl_Frequency
perforated	past participle	cl_PerforatingLesion
oto	prefix	cl_Ear
itis	suffix	cl_InflammationLesion
with	preposition	rel_Accompanying
purulent	adjective	cl_Pus
discharge	noun	cl_Discharge

Figure 1: Word recognition applied to the term *"frequent perforated otitis with purulent discharge"*

In a recent work [7], the authors experimented a semantic tagger for discharge letters in the digestive surgery sub-domain. They have compared tagging based either on a lexical or a semantic tagset applied to the same corpus of medical text and they have achieved in both situations a 98% disambiguation. Further work is expected to provide better results from semantic tagging, whereas the lexical tagging reaches inherent limitations. Semantic tagging is an important emerging technique. It is slightly dependent on the existence of a model of the domain. This aspect is dealt with in another paper by the authors [8].

A RULE-BASED PARSER

Parsers have been largely described in scientific literature about computational linguistics. In the medical community however, we are left with a gap preventing us from easy access to those techniques. The reasons for this gap are: the medical domain is relatively unknown to linguists; platform and implementation dependencies do exist; links from a scientific community to another are not well established; we deal with quite pragmatic (even non rational) situations when faced with a clinical setting; semantical approaches are necessary in more and more situations and domain model dependencies are not taken into account.

From the observation of medical language, we came to a simplifying hypothesis: patient description is mainly done through descriptive noun phrases and verbs do not vehicle much information. When verbs are present, they can, in most situations, be easily replaced by a related noun. Therefore, we can limit

ourselves to non verbal sentences (except past and present participles) and we decided – at least temporarily – to limit our experiments to such phrases.

Indeed, noun phrases form the large majority of terms in classifications and nomenclatures like ICD, SNOMED International, READ classification, ICPC, etc. This is also the chosen approach for controlled vocabularies [9]. It should be understood that such a limitation is a working hypothesis. We do not foresee major problems to extend our approach to account for verbs in the future. Our current tools already recognize and analyze verbs in any form, but we know that this will increase the probability of ambiguous situations, and therefore the processing time.

We designed a rule-based parser close to the so-called shallow parsing by finite-state automaton as referenced in the scientific literature [10]. Our rule-based parser is based on a set of rules defining transformation actions on the initial list of items. Each rule may act on 2 or 3 consecutive items, making a compound item of them. A rule may be fired more than once on the given list. The order of the rules is strictly defined and applied. A successful parsing of a term is obtained when the initial list of items has been transformed to a single compound item. When left with more than one item at the end, we are faced with a partial parsing, which provides useful information for further treatment.

Let us examine an example, like the sentence given above. The first rule to be applied considers the prefix-suffix pair *ot–itis*. It will say that the suffix is modified by the prefix, meaning that the suffix is the main morpheme at this level. The modifier is represented by a lexical link of the type *hasPrefix*. Such a link is the temporary vehicle for a semantic resolution - depending on an underlying model - to a semantical link like *hasSpecificLocation*. The grouping of the two items makes a compound item of type noun. The result will be:

```
cl_InflammatoryLesion
   hasSpecificLocation cl_Ear
```

The next rule will be applied three times: it deals with the adjective-noun pairs. It will be processed from right to left (suitable for adjectives in English, but wrong in some other languages). It will sequentially group the pairs *purulent-discharge*, *perforated-itis* and *frequent-itis*. We freely interpret the lexical grouping noun-adjective by the link *hasModifier*. A part of the result is:

```
cl_InflammatoryLesion
   hasSpecificLocation cl_Ear
   hasModifier cl_Frequency
   hasModifier cl_PerforatingLesion
```

The next rule will consider the pair preposition-noun pair *with-discharge* giving a compound item of type prepositional group. This paves the way to the last rule in this example tailored to a noun-prep-group pairs like *itis-with*. The final result is:

```
cl_InflammatoryLesion
   hasSpecificLocation cl_Ear
   hasModifier cl_Frequency
   hasModifier cl_PerforatingLesion
   Accompagnying cl_Discharge
      HasModifier cl_Pus
```

We have fired 5 rules. We got a single compound item which is a necessary condition for successful parsing. We clearly discovered the focus of the sentence. The final result is a form of knowledge representation of the initial term, supposedly independent of the source language. This means that the same result would be obtained when working either in different languages or with variable forms of the same term.

The order of application of the rules is very important. One can see from this example that the pair *perforated-ot* is wrong: not the *ear* is *perforated*, but *perforated* is a qualifier of this specific *inflammation*. Because the prefix-suffix rule is applied before the noun-adjective rule, the result is correct. Indeed, the prefix-suffix rule has to decide which is the main word from the two, and this problem has not yet been totally solved! The direction of application of the rules plays also a role: right to left is valid for adjective in English whereas left to right would fail.

A difficult problem is the mapping from the lexical links, as discovered by the rule, to some more meaningful links expressing the semantics of the sentence. Different solutions are possible: first, some words point directly to some semantic links, like *with* points to rel_*Accompagnying* in Figure 1 (though in some situations *with* may be ambiguous); second, we can rely on an underlying model of the domain – like GALEN [11] – from which we may infer that a body lesion like *inflammation* and a body part like *ear* are usually linked by the *hasSpecificLocation* link. The advantages, promises and pitfalls of modeling has been covered in other articles, which were at the center of a recent working group on this topic [12].

A TYPOLOGY OF RULES

Our current implementation of the morpho-semantem parser in English and French accepts 6 different types of rules. They are: 1) direct dependence of two

consecutive items; 2) enumeration of 2 items without specific order; 3) enumeration of 2 items governed by a third item (like conjunction or comma); 4) inverted dependence of two consecutive items; 5) other inverted dependence of two consecutive items; 6) direct dependence of two consecutive items with an intermediate item.

Our set of roughly 50 rules per language is consistent with a good rate of success (above 90% of all sentences correctly parsed) as long as one is limited to noun phrases without conjunction of coordination or commas. When coordination and commas are present in enumeration the rate of success may decrease to below 80%. Then, the idiosyncrasies and intricacies of medical jargon may lead to even lower values in the presence of common medical texts. However, we can foresee interactive useful applications in the coming years for clinical settings.

The structure of a rule is basically the following: each rule is defined by two lexical categories, which are adjacent in a sentence. An intermediate word can be specified in addition. From those two starting categories, the rule points to a final category which will be the result of firing the rule and available for further processing of other rules. In other words, each rule transforms two (or three) lexical items into one.

Typically, consider the *noun-adjective* structure which will give a noun (or a noun group but, from there on, acting as a noun). Another common situation is the *noun-of-noun* structure in the presence of a noun complement. The *structure proper name-proper name* is less frequent but often used for denomination of diseases or surgical deeds. Other situations are linked to articles, adverbs, preposition, etc.

Different parameters are allowed to increase the power of rules to cover any situation. The *direction* may be left to right or the reverse. For example, the *noun-adjective* structure is processed from right to left for English adjectives, because this allows the processing in one pass of multiple adjectives. *Lexical coherence* is another feature important in French and German, but not really in English. In these former languages accordance in gender and number is explicit in adjectives and articles. The rule may have to verify this feature when requested.

Last point of capital interest is the order of rules. Rules are fired in a specified order, the most specific rules being considered first. Changing the order may change the result. For example with a noun complement where each noun is accompanied by an adjective, the noun-adjective rule is applied first. Tuning the order of rules is considered as difficult, because some change may destroy what was already satisfactory with other sentences.

ADVANTAGES AND LIMITATIONS

Morpho-semantic parsing has shown its practical potential to analyze medical phrases. It is a variation of existing techniques, and before spending resources on further developments, an impartial look at its advantages and current limitations is necessary.

Advantages

- Morpho-semantic parsing is "natural" and closer to the final grain of knowledge representation;
- Morpho-semantic parsing is especially valuable in the medical domain due to the way it composes new terms from known ones;
- Lexicons can be largely reduced in size: from one third for a 20000 word lexicon to more than one half for larger lexicons;
- Morpho-semantic parsing has the ability to cope with new compound words which are not present in the lexicon;
- Processing time, being somewhat dependent on lexicon size, may be improved in the presence of shorter lexicons;
- Morpho-semantic parsing is robust: even if unsuccessful, this technique always leaves some partial practical result;
- The rule-based approach is a simple heuristic approach, which can be mastered by any native speaker of the language;
- Multilingual lexicons are easy to build due to similarities between Western languages;
- The rule-based approach may be tailored to the medical domain and the so-called medical jargon with many short-cuts from scholar language;
- Morpho-semantic parsing has already proved its value for simple tasks like indexing and retrieval;

Limitations

- Morpho-semantic parsing is not well known to computer linguists and available tools are limited;
- The disambiguation process is a necessary step and the availability of a tagger or similar technique is a must;
- Multilingual morpho-semantem lexicons are not yet available;
- The rule-based approach necessitates the fine tuning of rules which is a time-consuming task;
- The rule-based approach may have to be upgraded to more complicated types of rules in

order to cope with unexpected situations unsolved with existing types of rules;

- Morpho-semantic parsing cannot be based only on lexical techniques and the need for an underlying model of the domain is important; the dependency on the existence of such a model is heavy;
- The proper handling of multiword expressions which are not constant (having flexions or intermediate optional words) is not implemented;

CONCLUSION

The morpho-semantem story is a long one, starting already in the seventies, but it did not have much implementation success at the time or even later. This is probably due to the lack of sufficient computer power on the desktop, which is no longer true. This explains the fresh interest and renewal of such a technique. This paper clearly advocates in favor of further developments in this direction.

The list of advantages and limitations is quite impressive. More work on the subject should decrease the weight of the limitation side, giving a speculative trend towards the other side of the balance. The industrial development of an intelligent medical editor with full text retrieval is at hand. Due to the robustness of the technique and the increasing power of desktop computers, it should be available on a low cost platform. It opens the way to future intelligent editors as addressed by the authors [13].

Finally, during our present experiment, we found totally ambiguous terms, at least at the language level, and we are in a documented position to urge people designing classification and nomenclatures to use Natural Language Processing tools for their future versions [14]. This will increase the precision and usefulness of their work.

References

1. RH Baud, C Lovis, A-R Rassinoux, J-R Scherrer. Morpho-Semantic Parsing of Medical Expressions. AMIA Annual Fall Symposium (formerly SCAMC) 1998, C Chute Ed's, JAMIA suppl 1998, p 760-764.
2. AW Pratt. On the Matter of Medical Linguistic. Proceedings MEDINFO'77, Shires DB & al eds, 1977 IMIA, pp 223-224.
3. MG Pacak, L Cousineau, W White. The Segmentation Approach to Dictionary Construction. Annual Meeting of the Association of Canadian Pathologists, Sherbrooke, Canada, June 1972.
4. LM Norton, MG Pacak. Morpho-semantic Analysis of Compound Word Forms denoting Surgical Procedures. Meth Inform Med., 22: 29-36, 1983.
5. A-M Rassinoux, RH Baud, P Ruch, J-M Rodrigues. Model-based Semantic Dictionaries for Medical Language Understanding. AMIA Annual Fall Symposium (formerly SCAMC) 1999, NM Lorenzi Ed's, JAMIA.
6. W Ceusters, P Spyns, G De Moor, W Martin. Syntactic-Semantic Tagging of Medical Texts: The Multi-TALE Project. IOS Press, 1998.
7. Ruch P, Bouillon P, Wagner JC, Baud RH, King M. MEDTAG project. Intermediate report, October 1998. Fond national de la recherche scientifique. OFES, Bern Switzerland. By the authors.
8. P Ruch, P Bouillon, RH. Baud, A-M Rassinoux, J-R Scherrer. MEDTAG: Tag-like Semantics for Medical Document Indexing. AMIA Annual Fall Symposium (formerly SCAMC) 1999, NM Lorenzi Ed's, JAMIA.
9. JJ Cimino, PD Clayton, G Hripcsak, SB Johnson. Knowledge-based Approaches to the Maintenance of a large controlled medical Terminology. J. Am. Med. Informatics Assoc. 1 (1994), pp 35-50.
10. S Aït-Mokhtar, J-P Chanod. Incremental Finite-State Parsing. Proceedings of ANLP'97, Washington, 1997, pp.72-79.
11. J Rogers, AL Rector. Terminological systems: Bridging the Generation Gap. Annual Fall Symposium of AMIA (formerly SCAMC), Hanley & Belfus Inc, 1997, pp 610-614.
12. CG Chute, RH Baud, JJ Cimino, VL Patel, AL Rector. Special Issue on Coding and Language Processing. Meth Inform Med, 1998; 37(4-5).
13. RH Baud. Present and Future Trends with NLP. International Journal of Medical Informatics 52 (1998), pp 133-139.
14. J-M Rodrigues, B Trombert-Paviot, RH Baud, JC Wagner, F Meusnier-Carriot. Galen-In-Use: Using Artificial Intelligence Terminology Tools to Improve the Linguistic Coherence of a National Coding System for Surgical Procedures. MEDINFO'98 procs. Amsterdam IOS Press, 1998:623-627.