Versatility of a Multilingual and Bi-directional Approach for Medical Language Processing

Anne-Marie Rassinoux, Ph.D., Christian Lovis, M.D.,
Robert H. Baud, Ph.D., Jean-Raoul Scherrer, M.D.
Medical Informatics Division, University Hospital of Geneva, Switzerland

At the dawn of the 21st century, we are experiencing an exponential growth of online information that is mostly textual, and that benefits from new electronic media, such as the World Wide Web (WWW), to be broadly diffused across borders. However, there is a gap to bridge between holding information and accessing in a relevant way the deep underlying knowledge. Multilingual natural language processing (NLP), once tuned, is certainly the best solution to cope with this era of textual information. This paper focuses on the lesson learned through the joint development of an analyzer and a generator of medical language, within a multilingual context. Concrete examples, derived from the efforts under way in the European GALEN-IN-USE project, illustrate the use of these linguistic tools for the handling of surgical procedures.

INTRODUCTION

The growing popularity of the Internet's World Wide Web (WWW), mostly due to its ease of use and its almost non-existent organizational rules, plays a key role in the diffusion of and accessibility to information. The proliferation of textual information on the Web is subject to new requirements[1] among which retrieval and comprehension are of concern. First, users seeking information need content-driven retrieval tools that return only relevant information. Second, once this information has been targeted, it must be directly understandable to the user. This implies overcoming the problem of communication across language barriers. The present need towards linguistic tools is obvious and is corroborated by the impressive booming of the linguistic industry during the last decade. Limited solutions are already available to the Internet community, as for example, the AltaVista's Translation Assistant[2] which "provides a tool to translate a grammatically correct document into something comprehensible, but not perfect". The proposed solution typically offers a word for word translation that discards many language intricacies (e.g. ambiguities, grammar rules, language styles, and jargons peculiar to a domain). Such concessions are not at all satisfactory for health care, which requires an accurate and error-free access to clinical information.

Communication in health care is facing a two-fold need. On the one hand, analyzing textual documents (such as discharge summaries or radiology reports) offers the potential to easily and rapidly access relevant structured medical information. On the other hand, generating textual reports and explanatory notes from structured data strengthens the effectiveness and expressiveness of communication, as well-written documents are easier to comprehend. Both approaches are vital in health care communication and are the subject of substantial and ongoing research and development[3,4] in the medical informatic community.

In this paper, we discuss the benefits of developing in parallel multilingual analysis and generation of medical texts, as well as the particularities relative to each approach. This has reinforced the necessity to manage, independently but in-line, linguistic and conceptual knowledge. These remarks are grounded from our involvement in the GALEN-IN-USE project for delivering linguistic tools that help to build and integrate multilingual surgical procedure classifications within a common framework.

BACKGROUND

It is widely recognized in the medical informatic community that medical language is a sublanguage whose interpretation requires substantial medical knowledge. In particular, special attention must be paid to dealing with specific medical jargon as well as common assumptions not explicitly mentioned in narratives. This has led to more and more concept-oriented approaches based on a model of medical concepts.

The GALEN Approach

The GALEN project aims at providing such a concept model through its Common Reference Model (CORE model) expressed in the GRAIL formalism.[5] It affords, in a language-independent manner, a high level ontology that organizes concepts and relationships in a multiple inheritance hierarchy, together with formal subsumption and multi-level sanctioning to constrain composition of sensible concepts.

The current phase of the GALEN project, renamed GALEN-IN-USE, strives to assist in the collaborative construction and maintenance of surgical procedure classifications.[6] Compositional descriptions of surgical procedures in GRAIL are based on an adaptation and extension of the pre-standard established by CEN ENV 1828. This allows the assertion of more or less complex surgical procedures, as one can be defined as a combination of several other nested procedures. Moreover, each procedure is identifiable through a surgical action or deed (e.g. *excision*, *evacuation*, *implantation*) that acts on some anatomical site or topography (e.g. *base of skull* or *lesion of the cranial cavity*). Other possible characteristics are instrumentation and device (e.g. *endoscope*, *pressure captor*), approach (e.g. *open*, *closed*, *percutaneous*), intention (e.g. *diagnostic* or *therapeutic*), extent (e.g. *partial* or *total*).

French rubric 099: "Exérèse d'une lésion du foramen magnum avec déroutement de l'artère vertébrale, par craniotomie"

```
[[SurgicalDeed]-
  (isMainlyCharacterisedBy)->[performance]-
    (isEnactmentOf)->[[Excising]-
      (playsClinicalRole)->[SurgicalRole]\]-
        (actsSpecificallyOn)->[PathologicalBodyStructure]-
          (LocativeAttribute)->[ForamenOccipitaleMagnum]\
        (hasSpecificSubprocess)->[SurgicalApproaching]-
          (hasSpecificSubprocess)->[[Incising]-
            (playsClinicalRole)->[SurgicalRole]\]-
              (actsSpecificallyOn)->[Skull]\\\\
  (isCharacterisedBy)->[performance]-
    (isEnactmentOf)->[[Moving]-
      (playsClinicalRole)->[SurgicalRole]\]-
        (actsSpecificallyOn)->[VertebralArtery]\\\].
```

Figure 1 – A surgical procedure expressing an *"excision of a lesion of the foramen magnum with deviation of the vertebral artery, by craniotomy"*, and its internal representation in CG

Figure 1 gives an example of an original rubric belonging to the neurology section of the new French national catalogue of procedures NCAM (Nomenclature Commune des Actes Médicaux). The modeling result in GALEN is depicted in the conceptual graph (CG) formalism,[7] which is directly inferred from GRAIL through a Definite Clause Grammar (DCG). This formalism fits the requirements for knowledge representation of medical language as it supports granular (detailed), generative, and compositional representation of medical information as well as inferences on it through well-defined formal operations.[8]

The Linguistic Tools

The internal representation in the GALEN model, which carries the deep meaning of the original rubric, looks quite complex (see Figure 1). This raises two issues. The first one is concerned with the production of this structured representation. At the present time, this step consists firstly in producing semi-automatically an intermediate representation (so-called a dissection) which is subsequently (nearly) automatically expanded into a GRAIL expression.[6] The second issue deals with the reading, by end-users, of such a complex structure in order to grasp its interpretation. It follows that authoring compositional representations directly in formalism such as GRAIL is difficult, time-consuming, and requires special model-interpretation skills. This is where linguistic tools can help.

Both an analyzer[9] (so-called RECIT) and a generator[10] have been developed as part of the GALEN project. The results obtained on the example given in Figure 1 are displayed in Figure 2.

Generation from the structured representation displayed in Figure 1:

- in **English**: "surgical *excision of a lesion of the foramen occipitale magnum by craniotomy with surgical movement of the vertebral artery*"
- in **French**: "*exérèse chirurgicale d'une lésion du foramen occipital magnum par craniotomie avec déroutement de l'artère vertébrale*"
- in **Italian**: "*asportazione chirurgica di una lesione del grande forame occipitale con craniotomia e con deviazione chirurgica dell'arteria vertebrale*"

Analysis of the French paraphrase for the rubric 099: "exérèse chirurgicale d'une lésion du foramen occipital magnum par incision du crâne, avec déroutement de l'artère vertébrale."

CG built by the RECIT analyzer:

```
[[SurgicalDeed]-
  (isMainlyCharacterisedBy)->[performance]-
    (isEnactmentOf)->[SurgicalExcising]-
      (actsSpecificallyOn)->[PathologicalBodyStructure]-
        (hasSpecificLocation)->[ForamenOccipitaleMagnum:#]\
      (hasSpecificSubprocess)->[SurgicalApproaching]-
        (hasSpecificSubprocess)->[SurgicalIncising]-
          (actsSpecificallyOn)->[Skull:#]\\\\
  (isCharacterisedBy)->[performance]-
    (isEnactmentOf)->[SurgicalMoving]-
      (actsSpecificallyOn)->[VertebralArtery:#]\\\].
```

Figure 2 – Results of generation and analysis for the example given in Figure 1

A few general remarks can be made here. First, multilingual generation ensures that information embedded into complex representation be directly accessible to end-users through language expressions

enunciated in his native, or at least known, language. Such a formulated interpretation constitutes a good means of validating the accurateness of the internal representation,[11] as it enforces the way information is nested in the representation (see the proximity of the information related to *excision* and *craniotomy* in the generated phrases shown in Figure 2). Second, the analysis yields a conceptual representation that reflects, in line with the model sanctioning, the meaning embedded in the input sentence. Because of the ambiguity and implicitness of natural language, it is more reasonable to give as input to the analyzer, the natural language paraphrase, which is directly built by the domain expert who expresses what he believes the rubric means. Until now, such a paraphrase helped to author the intermediate representation of the corresponding rubric accurately and quickly. It follows that the representation built by the analyzer (see Figure 2) can be automatically compared with the initial internal representation in GALEN (see Figure 1) by applying formal conceptual operations, such as the projection operation.[8]

A BI-DIRECTIONAL APPROACH FOR NLP

Analysis and generation are commonly pointed out as being reverse processes. Indeed, the input of one is the output of the other, and vice-versa. The different steps involved during these two NLP tasks present valuable shared features, as roughly resumed in Figure 3.

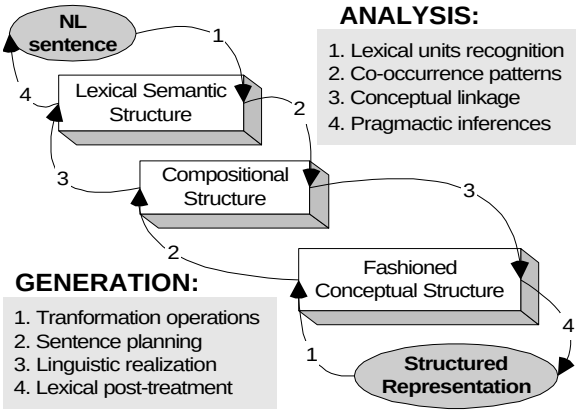

Figure 3 – Steps performed by NLP tools (presented here in sequential order for the sake of simplicity)

The complexity underlying each processing step together with the way knowledge is effectively used, differ with the NLP task considered. But it appears that a multilingual and bi-directional approach benefits from sharing substantial linguistic and conceptual knowledge, as explained in the following sub-sections that examine each step in turn.

Lexical Semantics: Tuning Forms and Meanings

Binding language words and conceptual structures is a key step for NLP, and is mainly realized through the handling of lexicons and semantic grammar. Lexicons are usually seen as the repository for 'computable' information about basic forms. These forms, being simple words, multi-word phrases, or smaller units such as prefixes and suffixes, can be seen as specific annotations for a concept (or more in the case of semantic ambiguity). Besides, semantic grammar intends to formalize the semantic relationships found in the domain through well-formed and sensible semantic co-occurrence patterns. These latter, usually denoting two concepts linked through a relationship, are called sensible statements in the GALEN's technology.[5] Their use by NLP tools has required the clarification of syntactic structures (such as adjectival or prepositional groups) that are commonly used in a specific language to support the expression of the corresponding relationship in a particular context.

NLP tools commonly use lexicons to mediate between language words and concepts. Indeed, analysis needs to retrieve every meaning (i.e. polysemy) associated with a recognized basic form in the lexicons. Therefore, the lexicons must have a good coverage of the domain treated. This point also ensures the expressiveness of the generated sentences as the generation mainly checks the lexicons for extracting an annotation for a concept according to a predefined syntactic category (mainly adjectives or nouns). Moreover as compound word forms are frequently used to denote surgical procedures,[12] their automatic handling (i.e. dealing with composition in generation and decomposition in analysis) considerably alleviates the size of lexicons. Moreover, NLP tools need semantic grammar for checking the semantic compatibility of two basic forms in analysis[9] and for choosing a syntax-based grammatical relation in generation[10] for expressing a specific pattern. Therefore, lexicons and semantic grammar appear as two shareable knowledge sources that, while being domain-dependent, take advantage of the multilingual approach to update in parallel their specific language-dependent parts.

Compositional Phase

The compositional phase is concerned with the way a clinical concept can be described, according to the others, in a specific context. For the generation (limited at the present time to the generation of noun phrases), this phase mainly deals with the sentence planning. This consists of delimiting the portion of the representation that will be worded in a whole

noun phrase as well as determining the appropriate use of conjunctions. For the analysis, the compositional phase consists of linking the various semantic fragments (i.e. parts of sentences) that were identified during the previous phase of lexical semantics, around one main concept that will be the focus of the structured representation built.

The common aim of this step, for both analysis and generation, is to identify the key relationships that serve to connect the different pieces of information and to correctly annotate them in the different languages treated. While the indented structure given as input to the generator makes explicit the way information is embedded, it has been necessary for analysis requirements to create a conceptual scheme for surgical deeds. The latter makes the manner that information is connected more explicit, thus allowing the right prepositional attachments to be performed around the main surgical procedure during the analysis process. For this, essentially two composite relationships *reld_isCharacterisedBy* and *reld_isNotCharacterisedBy* have been introduced for NLP needs in order to embody the performance and non-performance of surgical procedures (see Figure 1). These are annotated by syntactic structures in the various languages treated, as for instance the respective prepositions *with* and *without* in English. Such relationships have proven to be efficient in coping with the compositionality of surgical procedures whatever the NLP task considered.

Fashioning Conceptual Structure

The NLP step, dedicated to fashioning the structured representation, mostly deals with the granularity of the representation on the conceptual side and with the vocabulary conciseness on the linguistic side. The aim of the generation task is to produce phrases "as detailed as necessary but as concise as possible". This entails simplifying the structured representation in order to mask the specific modeling styles, which can subsequently corrupt the quality of the generated sentence by introducing idiosyncratic elements. However, the effects of modeling styles, reduced during generation, must be inversely restored during analysis. Indeed, the aim of the analysis task is to ensure the automatic construction of GRAIL representations that afford the same degree of detail as that actually performed in the GALEN model. This consists of expanding the representation built toward including implicit as well as pragmatic information.

The body of conceptual knowledge shared by analysis and generation during this phase is mainly constituted by conceptual definitions. These include: concept definitions (e.g. the concept *SurgicalMoving* is defined by *[Moving] - (playsClinicalRole) -> [SurgicalRole]*), relation definitions that are exclusively added to the GALEN model for NLP purposes (such as the two composite relationships previously described: *reld_isCharacterisedBy* and *reld_isNotCharacterizedBy*), and a few specific structures to manage conceptual behaviors such as the handling of the focus (see Figure 1 where the main procedure of the rubric, denoted by the composite concept *SurgicalExcising*, is embedded in the more general concept *SurgicalDeed*). Only the way such definitions are applied for a specific NLP task differs. Indeed, during the generation process these definitions are usually contracted, whereas they are usually expanded during the analysis process. These formal operations, fully explained elsewhere[8], allow for dealing with the conciseness of the vocabulary (concept definitions), with the modeling style (relation definitions), and with the focus of the representation that constitutes the central wording upon which the rest of the sentence is built. It follows that once created, a definition becomes available to any NLP task.

RESULTS

A large-scale experiment of generating natural language phrases for surgical procedures modeled in GRAIL has been achieved for the French NCAM classification. More than 700 rubrics belonging to the urology, gynecology, vascular surgery, and recently, neurology sections have been modeled first in GRAIL and then regenerated into French, English, and to a minor extent Italian. Results obtained from the last sample treated (composed of 100 neurology rubrics), yielded 11% of errors identified by experts of the domain. 5% resulted from a wrong dissection, 5% occurred during the expansion of the dissection into the GRAIL representation and only 1% was due to the generation process. These results show that the generator is now well tuned to the modeling style adopted for the handling of surgical procedures. The efforts needed to cover additional sections of this domain mainly result in adding annotations for the new concepts introduced. Besides, the analysis of paraphrases, currently written in English or French, has just started for the same sample of 100 neurology rubrics. The initial results are encouraging (see Figure 2) and the use of such a tool, to help build the internal representation in GRAIL for surgical rubrics, appears conceivable. However, interpreting any phrase in a given language would seem to be an

unreachable task as it would require a full coverage of lexicons as well as language structures.

DISCUSSION AND CONCLUSION

We have reported here on the benefits of sharing data during the analysis and generation of medical language, while being aware of the way such data is used for a particular NLP task. For this, it is important to continually manage with the refinement of the model, that is to say to be aware of new added details or implicit knowledge, as well as specific modeling styles introducing 'artefact' concepts to design a particular knowledge. The evolution of the concept model must also guide the linguistic tuning which is mainly concerned with the annotation process for concepts (i.e. lexicon adds) and for relationships (i.e. syntactic structure descriptions), in the various languages treated.

Working both with multiple languages (limited to European languages by the project scope) and with dual NLP tasks (analysis and generation) has stimulated the reasoning by analogy. This has strengthened the modularization of knowledge, thus facilitating the increase and maintenance of the knowledge base. Such comparative method was also feasible due to the fact that only latino-greek languages, which share common linguistic features, were considered.

Moreover, it is worth noting that successful multilingual communication does not rely on sentence-by-sentence linguistically correct translation but rather on a good understanding of the intended goal of communication that takes place in a specific context. Such a meaning expressed in the language-independent structured representation, acts as the interlingua. This allows the analyzer and generator to be used in sequence to produce purely conceptual translation. The resulting feedback loop between natural language phrases and conceptual representation also constitutes a valuable tool for assessing the outcomes in different languages as well as evaluating the accurateness of the compositional modeling.[11]

Finally, content-driven access for the recovery and extraction of relevant multilingual medical information would appear to be the best paradigm shift in health care to face the 21st century, already oriented towards the information era. This requires handling conceptual knowledge bases that, once bridged to linguistic knowledge, provide the foundation on which various NLP tasks can be developed.

Acknowledgments

The GALEN-IN-USE project is funded as part of Framework IV of the EC Healthcare Telematics research program. The Swiss part of the project is fully supported by the Swiss government (OFES - Office Fédéral de l'Education et de la Science).

References

1. McLeod SD, Gieser JP. Knowledge or noise. Scientific publication and the electronic journal. Arch Ophthalmol, 1996; 114(10): 1269-1270.
2. AltaVista and Systran. About AltaVista Translations. WWW page http://babelfish. altavista.digital.com/cgi-bin/translate?, 1997.
3. Spyns P. Natural Language Processing in Medicine: An Overview. Meth Inform Med, 1996; 35(4/5): 285-301.
4. Cawsey AJ, Webber BL, Jones RB. Natural Language Generation in Health Care. JAMIA, 1997; 4: 473-482.
5. Rector AL, Nowlan WA, Glowinski A. Goals for Concept Representation in the GALEN project. In: Safran C (Ed.). Proceedings of SCAMC'93. New York: McGraw-Hill, Inc. 1993: 414-418.
6. Rogers JE, Rector AL. Terminological Systems: Bridging the Generation Gap. In: Masys DR (Ed.). Proceedings of the 1997 AMIA Annual Fall Symposium. 1997: 610-614.
7. Sowa JF. Conceptual Structures: Information Processing in Mind and Machine. Reading, MA: Addison-Wesley Publishing Company, 1984.
8. Rassinoux A-M, Baud RH, Lovis C, Wagner JC, Scherrer J-R. Tuning up Conceptual Graph Representation for Multilingual Natural Language Processing in Medicine. Accepted for ICCS'98. Montpellier, France, August 10-12, 1998.
9. Rassinoux A-M, Wagner JC, Lovis C, et al. Analysis of Medical Texts Based on a Sound Medical Model. In: Gardner RM (Ed.). Proceedings of SCAMC'95. Philadelphia: Hanley&Belfus, Inc., 1995: 27-31.
10. Wagner JC, Solomon WD, Michel P-A et al. Multilingual Natural Language Generation as Part of a Medical Terminology Server. In: Greenes RA et al. (Eds.). Proceedings of MEDINFO'95. North-Holland: HC&CC, INC., 1995: 100-104.
11. Baud RH, Rodrigues J-M, Wagner JC et al. Validation of Concept Representation Using Natural Language Generation. In: Masys DR (Ed.). Proceedings of the 1997 AMIA Annual Fall Symposium. 1997: 841.
12. Norton LM, Pacak MG. Morphosemantic Analysis of Compound Word Forms Denoting Surgical Procedures. Meth Inform Med, 1983; 22(1): 29-36.