NLP Techniques associated with the OpenGalen Ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent Linguistic and Logical constructs

M Biczyk do Amaral [1,2], A Roberts [2], AL Rector [2]
[1] Sao Paulo University, Medical School Hospital, Brazil
[2] Medical Informatics Group, Dept Computer Science, University of Manchester, UK

This research project presents methodological and theoretical issues related to the inter-relationship between linguistic and conceptual semantics, analysing the results obtained by the application of a NLP parser to a set of radiology reports. Our objective is to define a technique for associating linguistic methods with domain specific ontologies for semi-automatic extraction of intermediate representation (IR) information formats and medical ontological knowledge from clinical texts. We have applied the Edinburgh LTG natural language parser to 2810 clinical narratives describing radiology procedures. In a second step, we have used medical expertise and ontology formalism for identification of semantic structures and abstraction of IR schemas related to the processed texts. These IR schemas are an association of linguistic and conceptual knowledge, based on their semantic contents. This methodology aims to contribute to the elaboration of models relating linguistic and logical constructs based on empirical data analysis. Advance in this field might lead to the development of computational techniques for automatic enrichment of medical ontologies from real clinical environments, using descriptive knowledge implicit in large text corpora sources.

INTRODUCTION

Since Wittgenstein pointed out the difference between concepts and language, the relationship between Natural Language (NL) and Knowledge Representation (KR) has proved elusive. The paradigm is often expressed in the well-known Ogden-Richards semiotics triangle. Following the theoretical work of Chomsky [1], Natural Language Processing (NLP) techniques have been used for various purposes. Computational linguistics has contributed to many techniques and methodologies for information analysis, storage and retrieval. However, it was not until recently that the interaction between language, logics and computation was addressed by formal analytical methods [2].

In this context, the development of medical terminologies is central to both NLP techniques and knowledge base (KB) construction in medicine [3]. Techniques associating NLP tools with medical terminologies for the purpose of data entry and automated text indexing have been described [4]. One important related work is MEDLee [5]. Applying NLP techniques to the extraction of knowledge from patient reports has proved to be a complex task. We have to consider not only the lexical and structural relations contained in the NL statements, but also the conceptual and pragmatic levels of medical knowledge that are embedded, at many levels, in the clinical narratives [6]. Central to addressing this problem is recognizing and representing equivalencies between linguistic structures expressed in written sentences, with conceptual structures associated with abstract logical models.

Research has reinforced the hypothesis that a model of medical concepts has to be closely related to the semantic components of a NLP parser [7]. Techniques for relating linguistic and conceptual semantics in medical terminologies, and their implications for NLP-based knowledge acquisition, have been developed [8].
An important approach for representing complex levels of conceptual knowledge has been pursued by groups developing medical ontologies [9]. The construction of such knowledge sources is a time consuming process, since most of the knowledge has to be obtained directly (manually) from medical experts. This problem could be solved by the discovery of techniques for building or automatically enriching medical ontologies, or the opposite, using ontologies for complementing the NL parsing process with domain knowledge.

The objective of our research project is to define a methodology for mapping linguistic structures into logical models. It aims at the abstraction of structural and conceptual semantic relationships between linguistic structures present in the clinical narratives and logical structures in the GALEN ontology.
Our research hypothesis is that the association of linguistic methods with domain specific ontologies, will enable the abstraction of structured intermediate representation (IR) schemas, containing conceptual

1067—5027/00/$5.00

and semantic patterns of the analysed knowledge domain. The set of obtained IR semantic structures might provide a basis for semi-automatic extraction of medical knowledge from clinical narratives.
Applying this NLP-Ontology methodology to the analysis of different sets of texts could contribute to a foundation for the interaction of conceptual and linguistic modelling in Medicine. Advance in this area might contribute to the development of better methods for semantic information storage and retrieval. In this paper, we describe methodological aspects of the technique we have developed.

MATERIALS and METHODS

Radiology Reports

Most clinical information organized in the patient records, either as paper or electronic, is written using free text. This is physicians' preferred way to report and record patient data. In Radiology, the narrative descriptions in the reports are important sources of clinical knowledge. In this project we have analysed 2810 medical narratives describing case reports of MR Scans, for the diagnoses of Acoustic Neuromas and Aneurysms. These reports are processed by NLP software, which tags and chunks parts of the narrative and organizes it as structured text.

The NLP Tool

The NLP tool used was the LT parser, which is a general purpose NLP program developed by the Edinburgh LTG (Language Technology Group - http://www.ltg.hcrc.ed.ac.uk/). It is a probabilistic part-of-speech tagger based on Hidden Markov Models using Maximum Entropy probability estimators. The LT parser comes with grammars to segment texts into paragraphs, segment paragraphs into words, recognise expressions, and mark-up expressions in texts. For our purposes, we have used two parts of the complete LT tool set: the LT POS and the LT Chunk modules. An additional chunker was developed by one of the authors (AR).
The LT POS part of speech tagger can handle plain ASCII text and SGML marked-up text. As a morphological classifier it uses a lexicon which is stored in a flat file and which can be easily extended to accommodate new words. LT POS achieves 96% to 98% accuracy when all the words are found in the lexicon and associated with their POS-classes by the morphological classifier. LT POS incorporates a part of speech guesser which employs a number of different guessing strategies. LT POS achieves 88-92% accuracy on unknown words. LT Chunk is a syntactic chunker or partial parser. It uses the part-of-speech information provided by a tagger and employs mildly context-sensitive grammars to detect boundaries of syntactic groups. The chunker leaves all previously added information in the text and creates a structural element, which includes the words of the chunk. It is capable of recognizing boundaries of simple noun and verb groups.

GALEN Ontology

The heart of the GALEN project is the *Open*GALEN concept reference (CORE) model of medical concepts which serves as the inter-lingua in which the concepts used in the medical record or referred to by coding systems or nomenclatures are represented. The *Open*GALEN model contains a well defined set of relationships between medical concepts based on description logic (DL) theories of generation and subsumption of composite concepts[9]. GALEN follows in a tradition of work on semantic networks and description logics first made explicit by Brachman[10]. A model in the GRAIL kernel consists of a network of nodes, called entities, and arcs connecting entities, called attributes. GRAIL statements consist of two entities linked by an attribute. The GRAIL kernel provides the rules for combining existing entities into new entities based on sanctions expressed by expressions in the model. The GRAIL language is like an assembly language because it is based on a logical formalism. In order to converge the linguistic semantic structure with the conceptual semantic structure of the GRAIL rules we can use the GALEN IR frameworks.

GALEN Ontology Intermediate Representation

The use of description logic techniques in GALEN allows a high level of detail and structure. This is necessary if GALEN is to meet its requirements. However, the detail and additional complexity makes it harder to meet other requirements, such as authoring, and the targeting of GALEN for language analysis. The GALEN Intermediate Representation provides a layer of abstraction, which reconciles these different requirements [11]. It acts as a high level language to the GRAIL assembly language. The discovery of ways to converge the equivalencies of the linguistic semantic IR construct and the GALEN conceptual IR constructs is central to this methodology.

Identification of IR Structures: Sets of Linguistic and Conceptual Patterns

The identification of the various levels and types of "structures" implicitly represented in the electronic medical record documents used by physicians during their clinical practice can be obtained by the

application of two processes that are called "segmentation" and "categorization" in the field of computational linguistics. Segmentation and categorization corresponds to De Sausurre's syntagmatic and paradigmatic linguistic division.

The segmentation and categorization of the radiology reports can be done at all possible levels in order to obtain a comprehensive and systematic set of structured component parts. First, we have to subdivide the reports into their internal main components, such as the procedure, the technique used, the anatomic location of the procedure, the description of the findings, the conclusion or final diagnosis, etc. The next step is to repeat the process for each statement of the reports. It is possible to perform this reasoning up to the atomic level of words or further into sememas ('atomic' semantic units). This will allow us to fragment the reports' internal structure and to see different levels of the hierarchy of linguistic and conceptual constructs present in the reports.

The identification of these semantic constructs and their inter-relationships is central to implementing IR schemas. This is the mechanism that will enable mapping from the linguistic structures found in medical texts and the logical structures responsible for representing medical knowledge in the ontology. What is required is a method to systematically and formally carry out the following transformation:

[Linguistic Construct]	⇔	[Conceptual Construct]
• language-dependent		language-independent
• linguistic sequences		logical relations
• grammar		ontology

This "equivalence mapping" may be implemented through the use of IR constructs that merge both types of patterns using the semantic information contents into a unique or unified schema.

PRELIMINARY RESULTS

The first level analysis for structuring the texts is the partitioning of the reports into the main components that describes its general structure. The empirical analysis of the reports showed that the medical narratives used for this procedure presents a quite organized macro-level structured sequence of what can be called the general framework of the report. It is a framework containing a regular set of classes for describing the performed clinical procedure, and other relevant clinical data, like diagnosis and anatomic location.

Structure of the Case Reports

The reports' general framework is the first 'structure' of the texts that we can use for identifying a conceptual model of clinical knowledge. The radiologists use this general framework to describe, using natural language, the facts associated with the realization of the procedure. Inside the general framework of the case reports, there is a multitude of possible NL descriptions of the findings and other clinically relevant information. However, many of these variations of NL descriptions are related to this conceptual or semantic macro model. The identified framework is the first structure that we used to relate the texts to a logical model. See figure 1 below.

[[<Date-Time stamp slot>]]
[[< PROCEDURE slot >]][[<ANATOMY slot>]]
[[<TECHNIQUE slot>]][[TECHNIQUE descriptiom>]]
[[<FINDINGS slot>]]: [[< FINDINGS description>]]
[[<CONCLUSION slot>]]:[[<DIAGNOSIS description>]]
[[<Signature/ Names MDs slot >]]

Figure 1: General framework found after empirical analysis of radiology reports describing MR scans.

When we say 'logical model' we mean a formal and precise description using the rigorous knowledge representation schema of ontologies. In terms of the GALEN ontology, it could either be a GRAIL statement or an IR logical structure.

Strings of syntactic patterns and chunks

The use of the LT parser for processing all the radiology reports produces a list of flat string sequences (part-of-speech tagged texts). Examples of strings of parts of speech tags are listed below:
CR1) nnp nnp nnp nn nnp nnp nnp jj nn nnp
CR2) nnp nnp nnp nn nnp nnp nnp jj nn nn nn
CRn) nnp nnp nnp nn vb nnp nnp jj ...
where CR means case report, nnp is a noun; jj is a adjective; vb is a verb, etc.

The next step is the application of the NLP parser to organize the structure of statements based on co-occurrence of word classes. This is done in part by the chunker. Co-occurrence of classes helps in the identification of other types of linguistic structures in the statements. The use of the chunker permits the organization of sequences of tagged words into phrases: bigger blocks composed of specific word sequences.

The importance of the chunker is the separation of phrases that can help in the identification of possible medical classes. In order to improve the level of

linguistic knowledge we have obtained from the analysed texts, it is necessary to add semantic labels.

Identification of Semantic and Linguistic Patterns

The next step towards the abstraction of semantic templates for mapping between linguistic and logical structures is the identification of semantic classes. The semantic blocks that compose the lexical semantic patterns are identified by classifying the strings with labels related to their classes. In our case, the semantic blocks are the ones describing medical classes. These classes are called concepts in GALEN. Some of the classes extracted empirically from the reports are listed below:
Semantic Classes/Categories: [PO] = Procedures;
[TC] = Technique; [AN] = Anatomy/Topography
[FI] = Findings; [DG] = Diagnoses/Diseases/Pathologies

These semantic categories are based on grouping or clustering words into meaningful classes used to describe the medical subdomain. See below:
[PO] ::= < MRI | USG | CT | X-Ray | ... >
[AN] ::= <brain | neck | abdomen | thorax...>
[FI] ::= <FINDINGS | ... >
[DG] ::= < Acoustic Neuroma | Angioma | Cyst | ...>

The next step is the abstraction of the semantic pattern that expresses the linguistic or logical relationship (links) between these classes. In the case reports that pattern is:
A Procedure [PO] which was performed at Anatomic_Location [AN], with technique [TC], showed results [FI], and the Diagnosis is [DG]

After identifying the main linguistic template, we looked at the GALEN ontology in order to define the equivalent logical construct.

Mapping the linguistic and conceptual levels

When analysing case reports, we can verify at least two levels for obtaining knowledge. One is related to the semantics contained in the overall structure of the report. The other is related to each individual sentence. In this phase of the project, we have focused only on the general framework. The advantage of using this approach is that it enables the abstraction of a simpler but highly relevant set of patterns. Particularly, this approach enables the identification of clinical pragmatics. The pattern is the type of knowledge that physicians look for when reading a radiology report in clinical practice. Therefore, it is an important piece of practical clinical knowledge. Furthermore, it clearly enables a straightforward association with the GALEN logical statements. The steps performed to converge the linguistic and the logical semantic patterns are:

- Identify the semantic chunks of the linguistic and conceptual structures
- Verify equivalencies of used symbols across different formalisms
- Map equivalent concepts from the different representations using domain knowledge

The repertoire of existing GRAIL formulas describing terminological knowledge and the relationships between medical concepts have to be consulted for pattern matching with the obtained linguistic semantic structures. Some examples of GRAIL statements found in the GALEN ontology that are applied to the analysed domain include:

Pathology which hasLocation Anatomy
Procedure which isPerfomed at Anatomy

Therefore, equivalence between linguistic semantic classes and the ontology logical categories are:

Linguistic Semantic Classes: [PO], [AN], ... [DG]
Ontology Categories: Procedure, Anatomy, Pathology

The logical interrelationship among those categories could be expressed as: *P (A(a)) => D(d)*

Meaning that: *a Procedure P, performed at Anatomical Location A(a) showed diagnosis or pathology D(d)*

Looking at the processed reports, this type of medical knowledge, related to the categories [PO], [AN] and [DG] are found in the head and tail chunk segments. The head, that is identified by the chunker as a single unit, contains the [PO] and [AN] semantic classes. The first chunk identified by the LT parser always has the format [PO_AN_TC]. The tail, which may be composed of single or multiple chunks, appears after the label "Conclusion" or "Comments".

With these equivalencies, we are able to define the mapping procedure between the linguistic and the conceptual levels. The linguistic semantic pattern is the IR model looking from the linguistic perspective. The conceptual semantic pattern is the IR model looking from the ontology perspective. When the linguistic semantic pattern matches the conceptual semantic pattern, then, we have found a common IR structure that merges linguistic and conceptual knowledge into a unified representation formalism. See below one example of a linguistic IR pattern and the equivalent conceptual IR pattern.

Linguistic IR Semantic Pattern:
[PO: {MRScan_NNP}]_[AN: {Brain_NNP}]_[TC:{axial turbo spin echo}]... []..._[DG: {AcousticNeuroma_NNP}]

Equivalent Conceptual IR schema:
MAIN magnetic ressonance scan
ACTS_ON brain
BY_TECHNIQUE axial turbo spin echo
WITH_RESULTS acoustic neuroma

This example shows the main semantic template of the reports, which frequently accounts for descriptions found in the reports. Other types of radiology procedures, like X-Rays, CTs, USG, also present a similar linguistic-conceptual pattern matching. Table 1 below shows some results related to the application of the methodology.

Table 1: Quantitative data related to the analysed reports and semantically labeled Noun Phrases (NP)

Semantic Classes of the Chunks	Aneurysm Reports: Total NP		Acoustic Neuroma: Total NP	
	Raw	Sort	Raw	Sort
All	10528	3552	67978	10013
Procedure	751	280	3615	299
Technique	1964	468	14045	862
Findings	6368	2519	40642	7778
Diagnosis	1445	773	9676	2429

DISCUSSION

A very important question was pointed out by Ceusters et al [8], regarding 'how do we go from conceptual ontologies to linguistic ontologies?'. Our research project showed a methodology that might contribute to advance in this direction. The application of modern computational methods has provided mechanisms for more precise analysis. Particularly, Bateman has shown that the semantics of a grammar used to describe a model play an important role in defining the ontology [12]. These results are in agreement with our methodology. The pivot point for the IR to be able to bridge the NL and the DL levels is the abstraction of the semantic relations and clinical knowledge contained in the texts [13]. There is a second level of segmentation that may be performed in order to further analyse the detailed NL sentences, that is to make a representation of each NL sentence and build the equivalent ontology IR. This has not been performed in the current phase of the project. We have defined some criteria necessary to identify the relevant semantics. We have used this criteria previously [14]. The first, from the linguistic side, is the successful identification of semantic chunks based on the parser's linguistic analysis. The second is the successful matching of chunks with categories found in the GALEN ontology. The last criteria is to verify that the structure is in accordance with the medical knowledge view of its contents.

CONCLUSION

The association of logics and linguistics could provide benefits, particularly in the improvement of methods for NLP and KR in medicine. Ontologies could be extended by the application of semi-automated methods to medical narratives. NLP techniques may use domain knowledge implicit in the logical formalism of the ontologies in order to provide support for robust NLP algorithms. The strengths of NLP techniques associated with domain ontologies, provide a good combination for identifying the semantic patterns of medical knowledge.

Acknowledgements
This project was supported by Sao Paulo State Research Foundation (FAPESP) grant #99/04291-0.

REFERENCES

1. Chomsky N. Knowledge of Language: Its Origins and Use. New York: Praeger, 1986.
2. Sowa J. Principles of Semantic Networks. Morgan Kaufmann, San Mateo CA, 1991
3. CG Chute, RH Buad, JJ Cimino, VL Patel, AL Rector (eds.). Meth Inform Med 1998; 37:311-575. Special Issue on Coding and NLP.
4. Sager N et al. NLP and the representation of clinical data. JAMIA 1994;1(2)142-60.
5. C Friedman et al. A general NLP for clinical radiology. JAMIA 1994;1(2)161-74
6. Scherrer JR. Concepts, knowledge, and language information systems. In [3] op cit.
7. Rassinoux AM et al. Modelling concepts in medicine for language understanding. In [3] op cit.
8. Ceusters W et al. The distinction between linguistic and conceptual semantics in medical terminology and its implication for NLP-based knowledge acquisiton. In [3] op cit.
9. Rector AL et al. Issues in Developing a Reusable Ontology for Medicine. In IEEE Trans Inform Tech Biomedicine; 2(4):229-242.
10. Brachman RJ, Levesque HJ. The tractability of subsumption in frame-based descriptiom languages . Proc AAAI-84, Austin-TX, 1984, pp.34-7.
11. Solomon D et al. How the GALEN IR reconciles internal complexity with users' requirements for appropriateness and simplicity. (submitted)
12. Bateman JA. Ontology Construction and Natural Language. Proc Int Formal Ontology and KR, 1993.
13. Baud RH, Lovis C, Rassinoux AM, Scherrer JR. Alternative ways for knowledge collection, indxing and robust language retrieval. In [3] op cit.
14. do Amaral MB, Satomura Y. Associating Semantic Grammars with the SNOMED. In Greenes et al (eds.). In MEDINFO95; p. 18-22.