Ceusters W, Spyns P, De Moor G. From Syntactic-semantic Tagging to Knowledge Discovery in Medical Texts. International Journal of Medical Informatics 1998;52:149-157.

Original article

From Syntactic-Semantic Tagging to Knowledge Discovery in Medical Texts

W. Ceusters [a, b] , P. Spyns [b] , G De Moor [b]

[a] Office Line Engineering NV, Het Moorhof, Hazenakkerstraat 20, B-9520 Zonnegem, Belgium
[b] RAMIT VZW, University Hospital, De Pintelaan 185, B-9000 Gent, Belgium

Abstract

In the GALEN project, the syntactic-semantic tagger MultiTALE is upgraded to extract knowledge from natural language surgical procedure expressions. In this paper, we describe the methodology applied and show that out of a randomly selected sample of such expressions coming from the procedure axis of Snomed International, 81% could be analysed correctly. The problems encountered fall in three different categories: unusual grammatical configurations within the Snomed terms, insufficient domain knowledge and different categorisation of concepts and semantic links in the domain and linguistic models used. It is concluded that the Multi-TALE system can be used to attach meaning to words that not have been encountered previously, but that an interface ontology mediating between domain models and linguistic models is needed to arrive at a higher level of independence from both particular languages and from particular domains.

Keywords: natural language processing; knowledge acquisition; concept representation

1. Introduction

The purpose of the GALEN project is to develop language independent concept representation systems as the foundations for the next generation of multilingual coding systems [1]. At the heart of the project is the development of a reference model for medical concepts (CORE) supported by a formal language for medical concept representation (GRAIL) [2]. A particular characteristic of the approach is the clear separation of the pure conceptual knowledge from other types of knowledge, including linguistic knowledge [3], in order to arrive in the future to application-independent medical terminologies [4]. Although on a theoretical basis the feasibility of these objectives is debatable [5], actual work within the GALEN-IN-USE project shows that on a relatively concise domain such as surgical procedures, distributed collaborative modelling can be achieved over linguistic borders. As could be expected, the process is however extremely slow. Formal "naming" and subcategorisation of new concepts at the one hand, and (in)consistent modelling of natural language expressions using the building blocks of the CORE that already are available, turn out to be the most frequent reasons for discussion.

```
(1) action repair noun     closed reduction > P1-10E30
(2)  -      -     prep     of
(3) do     path   sg       fracture of zygoma or zygomatic arch
(4)  -     path   sg            fracture of zygoma
(5)  -     path   sg               fracture > M-12000
(6)  -      -     prep             of
(7)  -     anat   sg               zygoma > T-11168
(8)  -      -     coor          or
(9)  -     anat   adjnoun       zygomatic arch > T-11167
```

Fig. 1: MultiTALE analysis of the sentence “closed reduction of fracture of zygoma or zygomatic arch”.

Given the promising results of the MultiTALE semantic tagger for neurosurgical procedure reports [6, 7, 8], it was investigated whether or not this manual modelling work could be speeded up by using MultiTALE as an automatic modelling device.

2. Material and methods

100 English surgical procedure expressions were randomly selected from the SNOMED International V3.1 procedure, excluding generic (codes P1-0xxxx) and anaesthetic (codes P1-Cxxxx) procedures. These expressions were then processed by the original MultiTALE tagger. The results were analysed to identify possible shortcomings at the level of the lexicon, the syntactic-semantic grammar and the desired format of the output, i.e. GALEN templates [9, 10]. Based on this analysis, a stepwise lingware refinement methodology was adopted, until a satisfactory number of expressions could correctly be processed.

The purpose of this study was then to investigate 1) whether the high level ontology of GALEN and the representation power of the GALEN surgical procedure templates were sufficiently elaborated for use in natural language understanding, 2) to identify what additional linguistic knowledge was needed to improve the results, and 3) to investigate whether the SNOMED expressions themselves could unambiguously be understood using the available conceptual and linguistic knowledge.

3. From MultiTALE to MultiTALE II

Multi-TALE is a syntactic-semantic tagger for English and Dutch neurosurgery reports [11]. The function of a tagger is to perform the first and essential pre-processing stage for any natural language processing application: the labelling of words with their grammatical categories, only taking into account a limited context. In addition to this syntactic labelling, the Multi-TALE tagger also provide semantic tags to words and word groups on the basis of the model for surgical procedures as described in CEN/ENV 1828:1995 [12].

Prior to any modification, MultiTALE analysed the expression

(s1) P1-11E52: *closed reduction of fracture of zygoma or zygomatic arch*

as an action of type repair which has as direct object a pathology, namely a fracture of zygoma or zygomatic arch (Fig 1). The semantic links discovered (action and do), as well as the semantic types (repair, path, anat) have their origin in CEN/ENV 1828:1995. In addition, for the individual concepts discovered, the SNOMED International code is given. Notice that the correct final results given in lines 1 and 3 of Fig 1 originate from an erroneous intermediate processing at lines 8 and 9 where the coordination is attributed at the wrong constituents. This is entirely due to the tagging nature of MultiTALE (as opposed to traditional parsers) according to which only the segmentation at the highest level matters.

```
np    {{Closed reduction} of {fracture of {zygoma or zygomatic arch}}}
np        { Closed reduction }
adj           Closed
noun          reduction
prep      of
np        { fracture of { zygoma or zygomatic arch } }
noun          fracture
prep          of
np            { zygoma or zygomatic arch }
noun            zygoma
conj            or
noun            zygomatic arch
```

Fig 2: MultiTALE II syntactic output of the expression "Closed reduction of fracture of zygoma or zygomatic arch"

```
RUBRIC "Closed reduction of fracture of zygoma or zygomatic arch"
MAIN reduction
   ACTS_ON fracture
         HAS_LOCATION zygoma / zygomatic_arch
   HAS_APPROACH closed
```

Fig. 3: MultiTALE II semantic analysis of the expression "Closed reduction of fracture of zygoma or zygomatic arch", presented in GALEN-template format.

With the objectives of GALEN in mind, this approach was no longer feasible as a more detailed analysis was required. MultiTALE was upgraded to MultiTALE II which produces the output of the same sentence (s1) as given in Fig 2 and Fig 3.

In order to achieve these results, the following changes to the original system were needed.

3.1 Implementation of a refined model for surgical procedures

According to CEN/ENV 1828:1995, a surgical procedure is conceptually composed of a *surgical deed* (i.e. deed that can be done by the operator to the patient's body during the surgical procedure) which is semantically linked to the concept fields *human anatomy*, *pathology* and *interventional equipment*. Potential semantic links are *direct object* (referring to that on which the surgical deed is carried out), the *indirect object* (referring to the (referring to the means with which the deed is site of the surgical deed) and the *means* carried out).

Although the standard was developed for the description of elementary surgical procedure expressions as they can be found in classification and coding systems, one of the hypotheses of the MultiTALE project was that the same structure could be used to represent the particular tasks described in full text reports. This turned out to be feasable, though not unproblematic. Especially the links *indirect object* and *direct object* were recognised to be underspecified for being useful within a natural language understanding environment, and lead to "non-monotonic like" semantic analyses. See for instance patterns such as:

(s2) Injection (deed) of antibiotic (direct object)

(s3) Injection (deed) of cyst (direct object)

(s4) Injection (deed) of antibiotic (direct object) in cyst (indirect object)

(s5) Irrigation (deed) of cyst (direct object) with antibiotic (means)

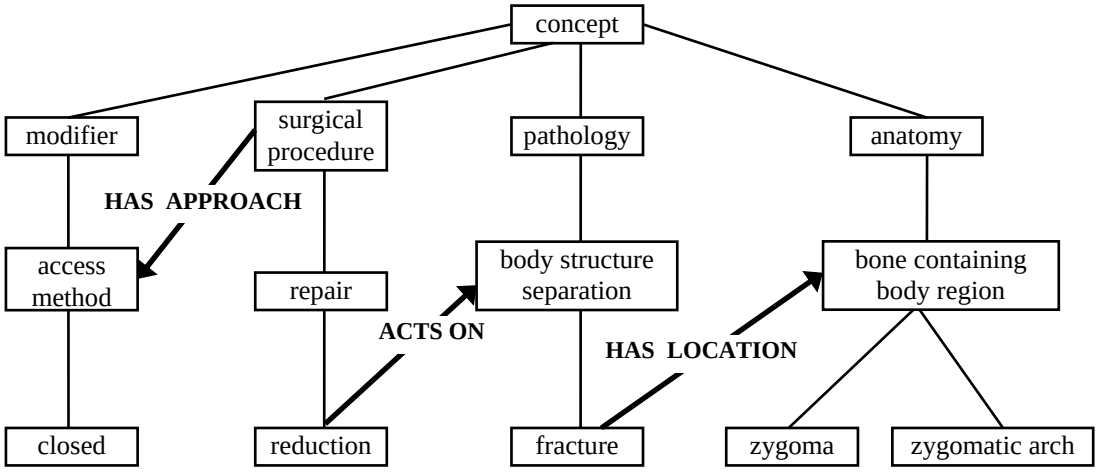

Fig 4: Relevant part of the concept hierarchy for the sentence *closed reduction of fracture of zygoma or zygomatic arch.*

For this reason, more refined links had to be foreseen such as *has_location*, *has_source*, *has_target*, *has_recipient*. As internally in MultiTALE II these links stand in a n-to-1 relationship to the original links of CEN/ENV 1828:1995, output can still be given according to the specifications of the ENV.

3.2 Implementation of a concept hierarchy

The MultiTALE tagger was directly based on the "flat" concept model of ENV 1828:1995, lexemes being encoded directly as *surgical_deed*, *anatomy*, *pathology* or *instrument*. To resolve certain linguistic ambiguities, a hierarchical model was needed. The relevant parts of the hierarchy needed to analyse the sentence of Fig 3, and the restrictional constraints on how some concepts may be linked, are outlined in Fig 4. However, in order not to duplicate the work of the modellers in the GALEN-IN-USE project, the conceptual model was not more enhanced than needed for an unambiguous interpretation of the expressions, leaving out the details required for generation purposes. In addition, only that part of the GALEN ontology that surfaces grammatically in the expressions, was incorporated [13]. A complete correpondance with the genuine GALEN model could however not be maintained as quite often the categorisation of concepts from a medical perspective does not follow the same semantic principles that underly sentence formation. Given the nature of surgical procedures where most often anatomical structures are being displaced or worked upon, we opted for a linguistically inspired categorisation where events are distinguished from entities, and further subdivided into states, acts, inchoatives and resultatives [14, pp 183-184]. For each of those, motional events are distinguished from non-motional events. As a consequence, procedures most often are motional acts involving thematic roles such as THEME, DESTINATION, SOURCE, LOCATION and PATH. Most of these thematic roles are then mapped to GALEN semantic links though not in a one-to-one fashion.

```
np        { Injection of xyz }                          RUBRIC "Injection of xyz"
noun        Injection                                   MAIN injection
prep        of                                          ACTS_ON xyz : chemical
noun        *xyz

              RUBRIC "Injection of xyz"
              MAIN injection
                    ACTS_ON chemical
                    HAS_DESTINATION xyz : body_part / body_region / pathology
```

Fig 5: Syntactic and semantic analysis of the sentence "injection of xyz".

3.3 Implementation of mechanisms for knowledge discovery

As MultiTALE II is designed to enrich the GALEN CORE and linguistic annotation modules (semi)automatically, mechanisms had to be foreseen for dealing with unknown words in the input. This was achieved using a bottom-up parsing strategy where both syntactic and semantic configurations limit each others possible interpretations. In Fig 5, the sentence

(s6) injection of xyz

(were xyz obviously is an unknown word) is analysed by MultiTALE II with one possible syntactic solution (xyz being a noun), and four possible semantic interpretations. First, xyz might be a body part, body region or pathology in which a not specified chemical is injected, as in

(s7) P1-10542: injection of ligament.

In these three cases, the HAS-DESTINATION semantic link applies. Next, xyz might be the chemical itself, with no destination specified, as in

(s8) P1-05027: injection of gas.

The various possibilities with respect to the meaning of xyz and the associated semantic links, are derived both from the internal concept hierarchy of MultiTALE-II (some relevant parts of which being shown in Fig 6) and from the possible syntactic configurations.

After dictionary lookup and grammatical analysis, the only possible syntactic configuration for the sentence is "noun-preposition-noun". The noun "injection" is found to have one possible meaning, namely an install procedure with as THEME a chemical. The preposition "of" can have several meanings, indicating a genitive semantic link (LOCATION, COMPONENT, ...), an objective link (THEME, PATIENT, EXPERIENCER, ...) or a receptive link (DESTINATION). In combination with "injection" only the objective and receptive interpretation are possible. The objective reading fixes the chemical as THEME, while in the receptive reading, body parts are candidate destinations. When finally the unknown word "xyz" is taken into consideration, it can unify with both the chemical and body-part readings.

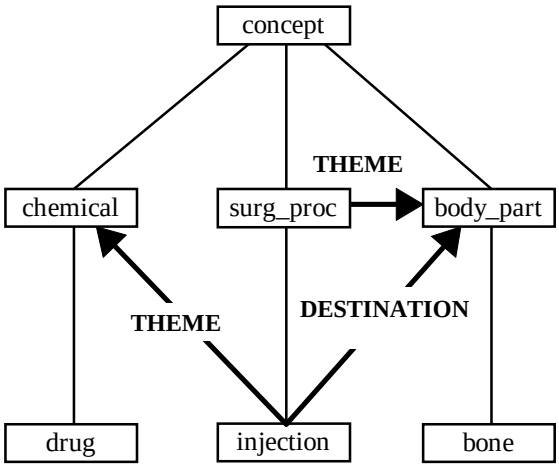

Fig 6. Relevant part of the concept hierarchy of MultiTALE-II with respect to the sentence "injection of xyz"

4. Results

Out of the 100 randomly selected expressions, 10 could not be processed by MultiTALE-II. For 7 of them, the required concepts or links were not yet available in the GALEN template-formalism, clearly a reason for failure outside the responsibility of MultiTALE. Of the remaining three, two showed peculiar (a)grammatical configurations:

(s9) P1-37865: Femoral-popliteal artery bypass graft with other than vein

(s10) P1-41335: Excision of benign lesion of scalp and neck, lesion diameter 3.1 to 4.0 cm)

while the other one contained deictic references and ellipsis, linguistic phenomena for which no grammar rules are currently implemented

(s11) P1-B9846: Bilateral repair of inguinal hernia, one direct and one indirect.

Of the 90 expressions that could be processed, 73 (81%) were analysed correctly, giving the only one possible interpretation, 58

of which by using exclusively the links foreseen in the GALEN template formalism, an intermediate representation developed in order not to confront the domain modelling experts with the complexity of the GRAIL language [15], while for the remaining 15, additional semantic links were introduced. It would have been possible to map these extra links to the "garbage"-link HAS_-OTHER_-FEATURE that is allowed in the templates, but we choose deliberately for not doing so in order to preserve the depth of the interpretation. 17 expressions led to multiple interpretations, 48 all together. Of those 48, 75% could be judged being correct.

5. Discussion

The results presented in this paper reflect not the final desired outcome of MultiTALE-II, but are rather to be seen as a first evaluation of the actual stage of the system, with further improvements in mind.

Ambiguities in the input phrases was the most important reason for multiple interpretations. E.g.

(s12) P1-A1122: Decompression of orbit only by transcranial approach

where "only" can refer to the orbit (nothing else being decompressed), to the decompression (nothing else than a decompression being done on the orbit), or to the approach (no other approach allowed for giving this code).

Coordination also led often to multiple interpretations, though the semantic constraints prevented all possible syntactic combinations, as can be seen in Fig 2, where syntactically a possible bracketing would have been: *{{Closed reduction} of {{fracture of zygoma} or zygomatic arch}}*. This possible syntactic solution is however not retained on semantic grounds.

Failure to reach an adequate interpretation was due to one of three reasons. For few sentences, the representational power of the GALEN-templates was not sufficient. It is for instance not yet possible to represent coordination amongst different semantic links that apply at the same time to one concept, e.g.

(s13) P1-2682B: repair of internal or complex fistula of trachea

where "internal" and "complex" specify two different features of "fistula". Also, the GALEN templates allow numbers to be linked to concepts using the HAS_NUMBER link, but quite often, an exact number cannot be deduced from the expression, as just a plural is given as in (s14) where one can only infer that there must be more than one adhesion.

(s14) P1-7AC34: Lysis of adhesions of spermatic cord

For some other sentences, specific surface linguistic constructs turned out to be problematic. E.g. in

(s15) P1-19B05: Primary suture of ruptured ligament of ankle, collateral

"Collateral" obviously specifies "ligament", but no grammar rule could yet be implemented in such a way that this sentence would be analyzed correctly without introducing erroneous output for other sentences such as in (s16).

(s16) P1-40141: Incision and drainage of hematoma, complicated.

Similar difficulties are caused by coordinated multiword units upon which ellipsis is applied, as in

(s17) P1-21A08: … rhinoplasty with lateral and alar cartilages ….

A third reason for incorrect results, is the lack of detailed anatomical knowledge such as the one required for correctly parsing (s18).

(s18) P1-17A26: Tenodesis for proximal interphalangeal finger joint stabilization

In this case, the system must know that "interphalangeal" refers to "joint" and not to "finger", in contrast with "abdominal wall mass" were "abdominal" refers to "wall".

6. Conclusion

The main conclusion of this work is that it is indeed feasible to develop a syntactic-

semantic parser that quite satisfactorily translates natural language expressions into a predefined formalism for further processing. However, in order to be able to extract new knowledge from texts, a certain amount of background knowledge, both conceptual and linguistic, must be available. All knowledge based approaches rely on an *ontology*, a more or less formal representation - to be used in computer systems - of what concepts exist in the world, and how they relate to one another. The GALEN ontology is viewed as a strictly language independent model of the world [16]. Meanwhile, the need for an ontology in natural language processing applications is generally accepted [17] as well. This is not to say that knowledge structuring based on a linguistic approach leads to the same result as when opting for a conceptual approach. A typical example is the ontological distinction between *nominal* and *natural kinds* [18], that in no language is grammaticalised just because the difference is pure definitional [19]. This again does not mean that such distinctions are not useful in a natural language processing applications. In MultiTALE-II for instance is the distinction between natural and nominal kinds used to analyse correctly expressions such as “capsulotomy of wrist” where at the surface “wrist” is connected to “capsulotomy” by an objective reading of “of”, while at deep level, the real THEME is the “capsule” that is located in the “wrist”. Also it has been stated that when s*ituated ontologies* (i.e. ontologies that are developed for solving particular problems in knowledge based applications [20]) have to operate in natural language processing applications, they are better suited to assist language understanding when the concepts and relationships they are built upon, are linguistically motivated [21].

A second conclusion of this work is that automated knowledge acquisition from nomenclatures such as SNOMED International, using natural language processing techniques, could be improved when controlled language principles would be applied for term formation in such nomenclatures. Such principles reduce ambiguity [22].

Based on these observations, future developments around Multi-TALE will concentrate on the design of an interface ontology mediating between situated ontologies and linguistic ontologies. This is expected to bring the Multi-TALE system to a sufficient level of independence from both particular languages and from particular domains.

7. References

[1] Rector AL, Nowlan WA, Glowinski A. Goals for Concept Representation in the GALEN project. In Safran C. (ed). SCAMC 93 Proceedings. New York: McGraw-Hill 1993, 414-418.

[2] Rector AL, Glowinski A, Nowlan WA, Rossi-Mori A. Medical concept models and medical records: an approach based on GALEN and PEN&PAD. Journal of the American Medical Informatics Association 1995, 2: 19-35.

[3] Rector AL, Nowlan WA, Kay S. Conceptual Knowledge: the core of medical information systems. In Lun KC, Degoulet P, Piemme TE, Rienhoff O (eds.). MEDINFO 92 Proceedings. Amsterdam: North - Holland 1992, 1420-1426.

[4] Rector AL. Compositional models of medical concepts: towards re-usable application independent medical terminologies. In Barahona P & Christensen JP (eds.) Knowledge and decisions in health telematics. Amsterdam: IOS Press 1994, 133-142.

[5] Ceusters W, Deville G, Buekens F. The chimera of purpose- and language-independent concept systems in healthcare. In Barahona P, Veloso M, Bryant J (eds.) MIE 94 Proceedings 1994, 208-212.

[6] Ceusters W, Deville G, De Moor G. Automated extraction of neurosurgical procedure expressions from full text reports: the Multi-TALE experience. In Brender J, Christensen JP, Scherrer J-R, McNair P (eds.) MIE 96 Proceedings. Amsterdam: IOS Press 1996, 154-158.

[7] Ceusters W, Deville G. A mixed syntactic-semantic grammar for the analysis of neurosurgical procedure reports: the Multi-TALE experience. In Sevens C, De Moor G (eds.) MIC’96 Proceedings, 1996, 59-68.

[8] Ceusters W, Lovis C, Rector A, Baud R. Natural language processing tools for the computerised patient record: present and future. In P. Waegemann (ed.) Toward an Electronic

Health Record Europe ’96 Proceedings, 1996:294-300.

[9] GALEN Consortium. Guidelines and Recipes for Completing templates. Internal document VUM02/96 version 1.0.

[10] GALEN Consortium. Links and Templates Summary. Internal document VUM/03/96 version 1.0.

[11] Ceusters W, Spyns P, De Moor G, Martin W (eds.) Syntactic-Semantic Tagging of Medical Texts: the Multi-TALE Project. Studies in Health Technologies and Informatics, IOS Press Amsterdam, 1998.

[12] CEN ENV 1828:1995. Medical Informatics - Structure for classification and coding of surgical procedures.

[13] Ceusters W, Buekens F, De Moor G, Waagmeester A. The Distinction between Linguistic and Conceptual Semantics in Medical Terminology and its Implications for NLP-Based Knowledge Acquisition. In: Proceedings of IMIA WG6 Conference on Natural Language and Medical Concept Representation. Jacksonville 19-22/01/97, 71-80.

[14] Frawley W. Linguistic Semantics (Lawrence Erlbaum Associates, Hillsdale, 1992).

[15] Rogers J, Solomon W, Rector A, Pole P, Zanstra P, and van der Haring E, Rubrics to Dissections to GRAIL to Classifications, in: C. Pappas, N. Maglaveras, J.-R. Scherrer, eds., Medical Informatics Europe ’97 (IOS Press, Amsterdam, 1997) 241 - 245.

[16] Rector AL, Rogers JE, Pole P. The GALEN High Level Ontology. In Brender J, Christensen JP, Scherrer J-R, McNair P (eds.) MIE 96 Proceedings. Amsterdam: IOS Press 1996, 174-178.

[17] Bateman JA. Ontology construction and natural language. In Proc. International Workshop on Formal Ontology. Padua, Italy, 1993, 83-93.

[18] Kripke S. Naming and Necessity. In Davidson D & Harman G (eds.) Semantics of natural language. Dordrecht: Reidel, 1972, 253-355.

[19] Welsh C. On the non-existence of natural kind terms as a linguistically relevant category. Paper presented at the Liguistic Society of America, New Orleans, LA, 1988.

[20] Mahesh K & Nirenburg S. A situated ontology for practical NLP. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95. Montreal, Canada, 1995.

[21] Deville G, Ceusters W. A multi-dimensional view on natural language modelling in medicine: identifying key-features for successful applications. Supplementary paper in Proceedings of the Third International Working Conference of IMIA WG6, Geneva, 1994.

[22] Ceusters W, Steurs F, Zanstra P, Van der Haring E, Rogers J. From a time standard for medical informatics to a controlled language for health. IMIA WG16 conference, Bermuda, 11-13/09/97 (unnumbered. Also to appear in International Journal of Medical Informatics).