Linking SweFN++ with Medical Resources, towards a MedFrameNet for Swedish

Dimitrios Kokkinakis

Department of Swedish Language, Språkbanken University of Gothenburg SE-405 30, Gothenburg, Sweden dimitrios.kokkinakis@svenska.gu.se

Abstract

In this pilot study we define and apply a methodology for building an event extraction system for the Swedish scientific medical and clinical language. Our aim is to find and describe linguistic expressions which refer to medical events, such as events related to diseases, symptoms and drug effects. In order to achieve this goal we have initiated actions that aim to extend and refine parts of the ongoing compilation of the Swedish FrameNet++ (SFN++). SFN++, as its English original predecessor, is grounded in Frame Semantics which provides a sound theoretical ground for modeling and linking linguistic structures encountered in general language and in specific domains (after specialization). Using such resource we have started to manually annotate domain texts for enriching SFN++ with authentic samples and for providing training data for automated event extraction techniques.

1 Introduction

In the clinical setting vast amounts of healthrelated data are constantly collected, while medical and biomedical scientific publications, in e.g. molecular biology, genetics, proteomics and other types of -omics, increase in a dramatic manner. These data are undoubtfully a valuable source of evidence-based research. However, to empower researchers to make highly efficient use of the resulting volume of literature and the knowledge that is encoded therein, the material must be better integrated and linked via effective automated processing. Tools have to be developed for the automatic processing of the textual content in a deeper, more semantically-oriented fashion having access to multilayered lexical and grammatical information. The goal is then to enable rapid, ef-

Maria Toporowska Gronostaj

Department of Swedish Language, Språkbanken University of Gothenburg SE-405 30, Gothenburg, Sweden maria.gronostaj@svenska.gu.se

fective and as far as possible accurate extraction of relationships, facts and events asserted and described in the data. Event extraction is understood here as an activity, that broadly follows the BioNLP 2009 shared task view (Kim et al., 2009), in which an event is considered to be an involvement of multiple entities in varying roles. The task is fundamental to the objective of Language Technology systems, such as Question-Answering and Information Extraction (IE), which have as their higher-level goal to identify instances of a particular class of events (or relationships) in a text and to extract their relevant arguments. We argue that such information has a direct correlation with FrameNet's semantic frames, since templates in the context of IE are frame-like structures with slots representing the event basic information. Our intention is to explore the applicability of SFN++ to the clinical and scientific medical domain in Swedish. Therefore, relevant domain specific entities are explicitly annotated by automatic indexing of the texts by the Swedish and English Medical Subject Headings thesauri (MeSH); cf. Kokkinakis (2009). Non-medical entities such as temporal expressions, locative expressions and personal characteristics such as gender and age are provided by an extended named entity recognition process (Kokkinakis, 2004). The partial syntactic analysis that follows is aware of the preprocessing steps and uses the background knowledge as features encoded in XML using the TIGER-XML format (Brants & Hansen, 2002).

2 Background

Methods employed in the extraction of events have generally involved two approaches. The first oneis based on manual annotation of events in domainspecific text samples in order to create training resources for processes that may *learn* to recognize events in new texts (Kim et al., 2008). The second is based on methods in which events are automatically acquired from unannotated texts; Nobata & Sekine (1999), in the sense that no manually preencoded training resources are used for producing the extraction patterns. Both methodologies have produced rapid advances in the robustness and applicability of IE. We believe that Frame Semantics (Fillmore, 1976) is a suitable resource, for the first type of method, and in our work we started to explore means for specializing and refining parts of the ongoing development of the SFN++ (Borin et al., 2009), on the medical domain. Our goal is to enrich domain corpora with layers of syntactic and semantic information providing relevant support for IE and text mining research in the field.

2.1 FrameNet

FrameNet <http://framenet.icsi.berkeley.edu> is based on the principles of Frame Semantics supported by corpus evidence. A semantic frame is a script-like structure of concepts, which are linked to the meanings of linguistic units and associated with a specific event or state. Each frame identifies a set of frame elements, which are frame specific semantic roles (e.g. participants or arguments). FN documents the range of semantic and syntactic combinatory possibilities of frame evoking lexical units, phrases and clauses. FN facilitates modeling the mapping of form and meaning within these structures in the medical discourse through manual annotation of example sentences and automatic summarization of the resulting annotations. A word can evoke a frame, and its syntactic dependents can fill the frame element slots (Dolbey et al., 2006). Moreover, since a single frame element can have different grammatical realizations it can enhance the investigation of combinatorial possibilities more precisely than other standard lexical resources such as WordNet (cf. Dolbey, 2009).

2.2 FrameNet and Information Extraction

IE is a technology that has a direct correlation with frame-like structures; since templates in the context of IE are frame-like structures with slots representing the event information. IE operates on specific domains, and the objective of IE systems is to identify instances of a particular class of events or relationships in a text and to extract the relevant arguments of the event or relationship. It has been stated Kilgarriff (1997) that the task requires the key terms of that domain, the "foreground lexicon", to be tightly bound to the domain vocabularies (e.g. ontology) and having wellarticulated meaning representations. According to this philosophy the foreground lexicon for a domain will generally contain:

- the key predicates for the domain (*trigger words*);
- the sets of lexical items which realize the predicate (*lexical units*);
- how they and their arguments relate to the IE system's output formalism (*core elements and valencies*);
- how their complements relate to the predicate's arguments (*non-core elements*).

3 Methodology

A subset of the original English FN already contains relevant frames with direct applicability to the medical domain, such as: Medical conditions, Experience bodily harm and Cure (see the Appendix for the SFN++ Cure frame; a snapshot from: <http://spraakbanken.gu.se/swefn/>); in the figure 'saldo' is the name of a freely available modern Swedish semantic and morphological resource intended for language technology applications; cf. Borin & Forsberg, 2009). We keep the English labels for each frame, while the content is manually adapted to Swedish. We start by identifying words or phrases that evoke frames and assigning the correct frame to them interactively using the SALTO tool (Burchardt et al., 2006). For each instance of a frame that has been found, we label the parts of the sentence that fill their semantic roles. Sentences that fulfill these criteria are selected from a large corpus after preprocessed by a number of pipelined tools including: multiword and idiom expression identification, part-of-speech tagging, lemmatization, named entity and terminology recognition, shallow parsing, using a cascaded parser (Abney, 1996) and XML formatting according to the TIGER-XML. A simplified example (i.e. The doctor treated her with cortisone) would at the end of the processing be labeled as:

FRAME[H Läkaren][LU behandlade][P henne][M med kortison]CUREHEALERLEXICAL UNITPATIENTMEDICATION

4 Conclusions and Further Research

Our work is a first attempt to get a whole picture of the requirements and difficulties for specializing SweFN++ to a domain and gaining experience on applying it to a sublanguage. Our goal for the near future is to aid the development of a fully automated event extraction system for the Swedish medical domain. This can be accomplished by annotating various types of data, in the medical sublanguage, and classifying text segments to the class type of the event mentioned. Then, by applying other means (e.g. pattern matching rules) we can extract the participants of the events and match to e.g. information extraction templates, depending on different applications and needs. Of course, there are several other issues that need to be worked on, such as devising ways to recognize negated and/or speculative language usage. Mapping medical frame elements onto corresponding concepts in a thesaurus-based lexicon turns a relatively little lexical resource into a more robust one and hence more useful for semi-automatic semantic annotation of corpora; cf. Baker et al., 2007. For annotating the Swedish corpus, we have used our intern thematically sorted lexicons with medical vocabulary and the Swedish data from MeSH.

Core FEs in FN	MESH thesaurus nodes
Ailment, Affliction	Disease
Body_parts	Anatomy
Medication	Chemicals and Drugs
Treatment	Analytical, Diagnostic
	and Therapeutic Tech-
	niques and Equipment
Patient	Person

Table 1. Example of mapping of core frame elements onto MeSH top nodes

The advantage of the pre-processing stage is very important and we believe that there is a feasible way to proceed in order to aid the annotation of large textual samples. Preliminary quantitative analysis of the examined instances has shown that many linguistically optional scheme elements need to be re-ranked whenever viewed from a medical pragmatic perspective. For example *Time*, *Measure* and *Method* provide relevant data for diagnosing patients' health condition. Another fact that might need special attention is the issue of tagging pronouns. It seems that these should not be tagged before anaphoric relations and their functional roles have been established. This is particularly important for distinguishing between patients and health care providers. Use of frame-semantic resources in general for language technology is evident. However, the effect of adding framesemantic information to LT applications has been rather low, our work intends to change this situation in the near future, getting insights into the deeper semantics of the domain events.

References

- Abney S. 1996. Partial Parsing via Finite-State Cascades. J. Nat. Lang. Eng., 2(4): 337-344.
- Baker C., Ellsworth M. and Erk. K. 2007. SemEval'07 task 19: frame semantic structure extraction. Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, Czech Republic
- Borin L. and Forsberg M. 2009. All in the family: A comparison of SALDO and WordNet. Nodalida Workshop: WordNets and other Lexical Semantic Resources between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. Odense.
- Borin L., Dannélls D., Forsberg M., Toporowska Gronostaj M. and Kokkinakis D. 2009. *Thinking Green: Toward Swedish FrameNet++*. FN Masterclass & Workshop. Milan, Italy.
- Brants S. and Hansen S. 2002. *Developments in the TIGER Annotation Scheme and their Realization in the Corpus.* 3rd Language Resources and Evaluation (LREC). Pp. 1643-1649 Las Palmas.
- Burchardt A., Erk K., Frank A., Kowalski A. and Pado S. 2006. SALTO – A Versatile Multi-Level Annotation Tool. 5th Language Resources and Evaluation (LREC). Genoa.
- Dolbey A., Ellsworth M. and Scheffczyk J. 2006. BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies. KR-MED: Bio. Ontology in Action. Maryland, USA.
- Dolbey A. 2009. *BioFrameNet, a FrameNet Extension to the Domain of Molecular Biology*. FRAMES AND CONSTRUCTIONS. A conference in honor of Charles J. Fillmore. U of California, Berkeley.
- Fillmore C. J. 1976. *Frame semantics and the nature of language*. NY Academy of Sciences: Conference on the Origin and Development of Lang. & Speech, Vol. 280: 20-32.
- Kilgarriff A. 1997. Foreground and Background Lexicons and Word Sense Disambiguation for Information Extraction. Proceedings of the Lexicon Driven Information Extraction. Frascati, Italy.

- Kim J-D., Ohta T., Pyysalo S., Kano Y. and Tsujii J. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. NAACL-HLT BioNLP-workshop. Boulder, Colorado.
- Kim J-D., Ohta T. and Tsujii J. 2008. Corpus Annotation for Mining Biomedical Events from Literature. *BMC Bioinformatics*, 8;9:10.

Appendix

Cure

- Kokkinakis D. (2004). *Reducing the Effect of Name Explosion*. Beyond Named Entity Recognition, Semantic labelling for NLP tasks. Workshop at the 4th LREC. Lissabon, Portugal.
- Kokkinakis D. (2009). Lexical granularity for automatic indexing and means to achieve it – the case of Swedish MeSH®. *Information Retrieval in Biomedicine: NLP for Knowledge Integration*. Prince V. & Roche M. (eds). pp. 11-37. IGI Global.

ram	Cure
domän	Med
semantisk typ	Cause_change_of_state
kärnelement	Affliction, Body_part, Healer, Medication, Patient, Treatment
periferielement	Degree, Duration, Manner, Motivation, Place, Purpose, Time, Type
exempel	Salvan lindrar även besvär som skavsår, sprickor på fingertopparna, stickor i fingrarna samt skärsår. Man kan behandla cancer med flera olika metoder. Läkaren opererade höger öga i stället för vänster. ST-läkaren behandlade henne med höga doser kortison. Salvan läker skrubbsår och brännsår. Genterapi botade dödssjuka i cancer. Transplantation kan ha botat hiv-smittad. Ljusterapi lindrar och förebygger nedstämdhet, ökar din energinivå och stärker ditt inre lugn.
sms	Type+Treatment, Body_part+Treatment, Medication+Treatment
sms-exempel	Type+LU_EX_ljus.behandling, röntgen.behandling, strål.behandling, värme.behandling Body_part+LU_EX_hjärn.operation, hjärt.operation Medication+LU_EX_kortison.behandling
saldo	vb: <u>avvänja1</u> <u>behandla2</u> <u>bota1</u> <u>hela1</u> <u>kurera1</u> <u>lindra1</u> <u>läka1</u> <u>läka2</u> <u>medicinera1</u> <u>operera1</u> <u>rehabilitera1</u> <u>vårda1</u> <u>återanpassa1</u> nn: <u>behandling2</u> <u>dialys1</u> <u>diatermi1</u> <u>hjärnoperation1</u> <u>hjärtoperation1</u> <u>huskur1</u> <u>läkning1</u> <u>kneipkur1</u> <u>kortvågsbehandling1</u> <u>kur1</u> <u>lindring1</u> <u>liusbehandling1</u> <u>lobotomi1</u> <u>medicinering1</u> <u>operation1</u> <u>radikaloperation1</u> <u>rutinoperation1</u> <u>rehabilitering1</u> <u>resektion1</u> <u>röntgenbehandling1</u> <u>skrapning1</u> <u>strålbehandling1</u> <u>värmebehandling1</u> <u>värmeterapi1</u> <u>återanpassning1</u>
saldo (nya)	
kommentar	Obs. behandling förekommer här som både Treatment och det rambärande lexemet, LU.; En sammanflätning (conflation) kan förekomma i objektpositionen mellan Patient och Affliction med vissa verb i den här semantiska gruppen, tex bota hiv-smittad/aids. (Se The Book, s.25 cure epileptic/epilepcy); (I Eng Cure-ramen tolkas 'affliction' som the injuries, disease, pain.);
skapad av	MTG
skapad	2010-01-30
modifierad	2010-02-27