User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-559

Factuality Levels of Diagnoses in Swedish Clinical Text

Sumithra VELUPILLAI [a,1], Hercules DALIANIS [a], Maria KVIST[a, b]
[a] *Dept. of Computer and Systems Sciences (DSV), Stockholm University, Forum 100, SE-164 40 Kista, Sweden*
[b] *Dept. of clinical immunology and transfusion medicine, Karolinska University Hospital, SE-171 76 Stockholm, Sweden*

Abstract. Different levels of knowledge certainty, or factuality levels, are expressed in clinical health record documentation. This information is currently not fully exploited, as the subtleties expressed in natural language cannot easily be machine analyzed. Extracting relevant information from knowledge-intensive resources such as electronic health records can be used for improving health care in general by e.g. building automated information access systems. We present an annotation model of six factuality levels linked to diagnoses in Swedish clinical assessments from an emergency ward. Our main findings are that overall agreement is fairly high (0.7/0.58 F-measure, 0.73/0.6 Cohen's κ, Intra/Inter). These distinctions are important for knowledge models, since only approx. 50% of the diagnoses are affirmed with certainty. Moreover, our results indicate that there are patterns inherent in the diagnosis expressions themselves conveying factuality levels, showing that certainty is not only dependent on context cues.

Keywords. Diagnosis reasoning, factuality levels, annotation, Swedish, clinical text, electronic health records.

1. Introduction

The process of diagnosing a patient is not trivial, and involves making decisions based on many diverse criteria. Clinicians are documenting reasoning processes and decisions in free-text, information that is currently not fully exploited for further knowledge management or research. Accurate and situation-specific information access is extremely important, especially in the clinical domain. This will provide clinicians with tools for information retrieval, using extracted information to produce relevant summaries, aggregating extracted information for knowledge discovery and further clinical research [1].

In order to create information access solutions that utilize the knowledge documented in free-text, it is necessary to be able to model subtleties expressed in natural language. One important aspect to consider is the level of certainty expressed in the reasoning and decision context. For instance, a likely scenario is the incorporation of a search engine in an electronic health record system, where clinicians can search for previous mentions of diagnoses for a particular patient. However, some of these diagnoses are written in a negated or speculative context, e.g. *this is definitely not*

[1] Corresponding author

diabetes or *angina pectoris cannot be excluded*. It is crucial that such distinctions are observed, as they convey different levels of knowledge certainty.

Research on modeling factuality levels, or degrees of certainty, in textual data, has increased in recent years. In the BioScope corpus [2], which contains biomedical texts, certainty levels are annotated at a sentence level, while negation and speculation cues are annotated at a token (word) level. In FactBank, factuality levels in newspaper articles are instead annotated on an event level [3]. In the clinical domain, agreement on probability expressions in radiology reports has been studied. Two studies analyzed phrases indicating different levels of certainty with respect to diagnoses [4, 5]. Both studies show that intermediate probabilities are more difficult to agree on while phrases indicating very high or low probabilities result in higher agreement. In automatic information retrieval settings, these issues have also been addressed in the research community lately. RadReportMiner [6] is a context-aware search engine, taking into account negations and uncertainties, achieving improved precision results (81%) compared to a generic search engine (27%).

In this paper, we present a model for annotating factuality distinctions in clinical documentation. Our aim is to develop automated systems that distinguish factuality levels of diagnoses in Swedish. Two clinicians annotate diagnoses in free-text entries for factuality levels. We analyze and evaluate the annotations with Intra- and Inter-Annotator Agreement (IAA). To our knowledge, this is the first attempt at modeling these distinctions and creating such a resource in Swedish.

2. Methods

Work process: we (1) assembled a list of diagnoses and created a resource for annotation, (2) developed annotation guidelines and annotated the created set, (3) evaluated Inter- and Intra-Annotator Agreement and did a qualitative analysis. We used the Knowtator plugin in the Protégé tool [7] for all annotation work. All documents were extracted randomly. Two senior physicians, A1 and A2, performed all annotation tasks, both accustomed to reading and writing medical records.

We extracted free-text entries from an emergency ward included in the Stockholm EPR Corpus [8]. Only entries documented under the category *Bedömning* (Assessment) were used in the annotation task. This field was chosen since it is the documentation entry containing most reasoning.

2.1. Creating a set of Documents Marked with Diagnoses

Instead of using diagnoses from Swedish medical terminology resources, we wanted to capture many diagnosis variants (e.g. inflections, misspellings, abbreviations). A collection of Swedish diagnoses was produced through a manual analysis of a subset of 150 assessment fields. A diagnosis was defined as a medical condition with a known cause, prognosis or treatment. All different variants and inflections of the same diagnosis expression were annotated.

A simple string matching procedure was employed to automatically mark diagnoses from the created diagnosis collection. A general language automatic

lemmatizer for Swedish[2] was used for capturing further inflections. Each diagnosis was marked with brackets, e.g. *Patient with <Diagnosis>diabetes</Diagnosis>*.

2.2. Annotation Classes and Guidelines

Factuality levels were modeled in two polarities: Positive and Negative. These were further graded: Certain, Probable or Possible. Each extracted diagnosis expression was annotated as belonging to one polarity and gradation, e.g. Certainly Positive, resulting in six annotation classes. Furthermore, the class Not Diagnosis was included for cases where the current context was not a diagnosis (e.g. infektion – short for clinic), and the class Other, for cases where e.g. the diagnosis referred to someone other than the patient, or where the annotator was uncertain. A first annotation task was performed in order to create detailed guidelines for the remaining task3.

2.3. Evaluation Metrics

The results were evaluated with IAA: F-measure, and Cohen's κ. IAA (Intra) results were measured on documents annotated twice by annotator A1, the second time in a new, randomized order. IAA (Inter) results were measured on documents annotated by two annotators; A1 and A2, treating A1 as the gold standard.

3. Results

In total, the number of annotated diagnosis instances was 2 182 (A1 vs A1) and 2 070 (A1 vs A2)4, extracted from 1 297 Assessment fields (approx. 51% of the total amount of Assessment fields). From the collection of 337 diagnoses, 227 were found.

3.1. Intra- and Inter-Annotator Agreement

A confusion matrix over the number of instances assigned to each class is shown in Table 1. *Certainly Positive* was in clear majority, almost 50% of the total number of instances. *Possibly Negative* and *Not Diagnosis* were very rare. The main discrepancies between the two annotators were in cases of assigning intermediate factuality levels. A1 generally assigned higher levels of factuality. Intra- and Inter-Annotator Agreement was very high for the majority class *Certainly Positive* (0.9 F-measure, respectively), while very low for *Possibly Negative* (0.35/0.03 F-measure, respectively), being a rare class. It is interesting to note that the classes *Not Diagnosis* and *Other*, both relatively rare, resulted in fairly high agreement results (0.82/0.62 and 0.69/0.65 F-measure, respectively). Overall IAA measured by Cohen's κ is: 0.73 (Intra), and 0.60 (Inter).

[2] http://www.cst.dk/online/lemmatiser/
[3] Annotation guidelines, including examples, can be found at http://www.dsv.su.se/hexanord/guidelines/ (guidelines_stockholm_epr_diagnosis_factuality_corpus.pdf)
[4] The discrepancy between the two sets is caused by mismatches and missed instances

Table 1. Confusion matrix, Intra- and Inter-Annotator Agreement.

	CP	PrP	PoP	PoN	PrN	CN	ND	O	Σ
CP Intra	**990**	78	4	0	3	4	2	19	1100
Inter	***834***	*59*	*7*	*0*	*4*	*5*	*1*	*20*	*930*
PrP Intra	20	**236**	55	1	1	0	1	0	314
Inter	*66*	***134***	*10*	*1*	*0*	*0*	*2*	*1*	*214*
PoP Intra	4	38	**127**	25	9	0	0	2	205
Inter	*11*	*149*	***180***	*41*	*45*	*1*	*1*	*10*	*438*
PoN Intra	0	0	6	**14**	7	1	0	1	29
Inter	*0*	*0*	*0*	***1***	*5*	*1*	*0*	*0*	*7*
PrN Intra	1	1	1	10	**118**	25	0	5	161
Inter	*0*	*0*	*0*	*2*	***35***	*18*	*0*	*1*	*56*
CN Intra	2	0	4	0	51	**195**	0	1	253
Inter	*2*	*0*	*0*	*4*	*99*	***193***	*1*	*3*	*302*
ND Intra	0	0	0	0	0	0	**26**	0	26
Inter	*13*	*5*	*3*	*2*	*1*	*3*	***30***	*4*	*61*
O Intra	8	1	4	1	7	0	8	**65**	94
Inter	*1*	*1*	*1*	*1*	*5*	*3*	*1*	***49***	*62*
Σ Intra	1025	354	201	51	196	225	37	93	**2182**
Inter	*927*	*348*	*201*	*52*	*194*	*223*	*36*	*88*	***2070***

Columns: A1, first annotation iteration. Rows: Intra: A1, second annotation iteration (same set randomized), Inter: A2. CP = Certainly Positive, PrP = Probably Positive, PoP = Possibly Positive, PoN = Possibly Negative, PrN = Probably Negative, CN = Certainly Negative, ND = Not Diagnosis, O = Other, **Σ** = Total

3.2. Qualitative Analysis

We also performed a manual, qualitative analysis of the resulting class assignments. We found that Certainly Positive dominated where a) diagnoses show overtly, e.g. skin diseases (eczema, urthicaria, skin infection) and general conditions (overweight, asystolia, fainting), or b) diagnosis was made by an apparatus (auricular fibrillation/ ECG). Probably Positive dominates for diagnoses with medical reasons for not securing certainty, e.g. virosis, gasthritis. Linguistic reasons seem to direct the following for some diagnoses: 1) an inverted pattern with a complementary vocabulary, e.g. ischemia (Certainly/Probably Negative in majority), heart attack or angina pectoris (Certainly/Probably Positive in majority), 2) a lack of negative annotation classes when normality was not expressed as negation (hypertension), 3) for lunginflammation (pneumonia), speculation was expressed in Swedish while we saw certainty expressed in Greek.

4. Discussion

In this study we present a model for knowledge certainty classification. This is used for the creation of an annotated set of Assessment entries from a Swedish emergency ward for factuality levels assigned to diagnoses. The model was functional and agreeable to the domain expert annotators. Our IAA results suggest that this model and resource can be used for developing automated systems. We also show, through a qualitative analysis, that factuality levels for different diagnoses are dependent on diagnosis type as well as inherent linguistic factors. This demonstrates that factuality and speculation in clinical text resides not only in linguistic context cues.

4.1. Limitations

The study design has some limitations that lowered the recall of diagnoses to be annotated. By employing a strict matching approach, yielding high precision, possible variants in form of misspellings, compounding and other formulations were missed. Fuzzier matching techniques could increase recall, at the cost of lower precision. The use of a limited list of diagnoses will inevitably result in a skewed distribution of diagnosis types. As a result, the model may not catch enough numbers and types of expressions of subtleties in conveying levels of factuality. How this in turn limits the created resources' ability to be used for machine learning is yet to be seen. The main limitation of this model for future work is the low numbers of annotations in some annotation classes. Intermediate probability assignments are clearly not self-evident (e.g. [4] and [5]). It can be argued that factuality levels *Possibly* and *Probably* may be fused, or even two *Possibly* classes, to lower the number of factuality levels, and increaseing training instances for machine-learning tasks. Such fusion was not agreeable to the involved physicians, as it would be a less accurate description of reality.

4.2. Significance of Study

Our results have important implications on the creation of intelligent information access from electronic health records. Without factuality analysis, uncertain or negated diagnoses would be identified as factual diagnoses. We have chosen a broad context-aware approach, in order to receive a wide perspective on how factuality levels are expressed concerning diagnoses. To our knowledge, no other studies have used a similar approach in this domain. Studies in the biomedical field (e.g. [3]) use hedge cues to detect uncertainty. We hope our approach will reveal inherent and previously unknown features that will aid in future machine-learning and text-mining studies.

Acknowledgments: This research has been carried out after approval from the Regional Ethical Review Board, Stockholm (Etikprövningsnämnden i Stockholm), permission no 2009/1742-31/5

References

[1] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JE. Extracting Information from Textual Documents in the Electronic Health Record, *IMIA Yearbook of Medical Informatics 2008* **47** Suppl. 1 (2008), 138–154.

[2] Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The Bioscope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes, *BMC Bioinformatics* **9(S-11)** (2008)

[3] Saurí R, Pustejovsky J. FactBank: a corpus annotated with event factuality, *Language Resources & Evaluation* **43** (2009), 227–268

[4] Khorasani R, Bates DW, Teeger S, Rothschild JM, Adams DF, Seltzer SE. Is Terminology Used Effectively to Convey Diagnostic Certainty in Radiology Reports?, *Academic Radiology* **10** (2003), 685–688.

[5] Hobby JL, Tom BDM, Todd C, Bearcroft PWP, Dixon AK. Communication of Doubt and Certainty in Radiology Reports, *The British Journal of Radiology* **73** (2000), 999–1001.

[6] Wu AS, Do BH, Kim J, Rubin DL. Evaluation of Negation and Uncertainty Detection and its Impact on Precision and Recall in Search, *Journal of Digital Imaging*

[7] Ogren P. *Knowtator: a Protégé plugin for annotated corpus construction*, in Proc. HLT-NAACL 2006, Morristown, NJ, USA, ACL, 2006, pp. 273–275

[8] Dalianis H, Hassel M, Velupillai S. *The Stockholm EPR Corpus – Characteristics and some Initial Findings*, in Proc. 14th ISHIMIR, Kalmar, Sweden, 2009.