Fine-grained Certainty Level Annotations Used for Coarser-grained E-health Scenarios

Certainty Classification of Diagnostic Statements in Swedish Clinical Text

Sumithra Velupillai[1] and Maria Kvist[1,2]

[1]Dept. of Computer and Systems Sciences (DSV)
Stockholm University, Forum 100, SE-164 40 Kista, Sweden
[2]Dept. of clinical immunology and transfusion medicine
Karolinska University Hospital, SE-171 76 Stockholm, Sweden
sumithra@dsv.su.se, maria.kvist@karolinska.se

Abstract. An important task in information access methods is distinguishing factual information from speculative or negated information. Fine-grained certainty levels of diagnostic statements in Swedish clinical text are annotated in a corpus from a medical university hospital. The annotation model has two polarities (positive and negative) and three certainty levels. However, there are many e-health scenarios where such fine-grained certainty levels are not practical for information extraction. Instead, more coarse-grained groups are needed. We present three scenarios: *adverse event surveillance*, *decision support alerts* and *automatic summaries* and collapse the fine-grained certainty level classifications into coarser-grained groups. We build automatic classifiers for each scenario and analyze the results quantitatively. Annotation discrepancies are analyzed qualitatively through manual corpus analysis. Our main findings are that it is feasible to use a corpus of fine-grained certainty level annotations to build classifiers for coarser-grained real-world scenarios: 0.89, 0.91 and 0.8 F-score (overall average).

Key words: Clinical documentation, Certainty level classification, Annotation granularity, Automatic Summary, Decision Support Alerts, Adverse Event Surveillance, E-health

1 Introduction

A challenging Natural Language Processing (NLP) task is to accurately extract relevant facts from clinical documentation. Speculative and negated information need to be distinguished from asserted information. Electronic health records are rich in factual and speculative opinions about a patient's clinical conditions, often expressed in free-text. This information is valuable for many e-health information access situations.

Certainty level classification in corpora is a growing research area in the domain of computational linguistics and information access, in particular for domain-specific purposes.

1.1 Related Work

In the interdisciplinary area of clinical natural language processing, several studies have targeted the issue of accurate information extraction by including negations and speculations in the information extraction model. In [1], assertion classification (present, absent or uncertain) is performed on medical problems. Rule-based and machine-learning techniques are used and compared. The machine-learning method, using features in a window of $\pm$ 4, outperforms the rule-based method. Contextual features, including negation, are used for classifying clinical conditions in [2]. In this study, uncertainties are, however, not modeled. The BioScope corpus contains annotations for negation and uncertainty [3] on a sentence level, with a subset of clinical radiology reports (the remaining corpus contains biomedical research articles and abstracts). The 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text [4] included a subtask for classifying assertion levels of medical problems. The top performing system on the assertion task obtained an F-score of 0.94 [5]. However, certainty levels are not modeled on a fine-grained level in these studies. In other domains, more fine-grained certainty levels are proposed, e.g. [6], [7] and [8]. The above-mentioned studies are performed on English.

1.2 Aim and Objective

In this work, we use a Swedish clinical corpus with diagnostic statements annotated at a fine-grained certainty level [9] to build coarser-grained classifications reflecting three e-health scenarios where this distinction differs for each scenario: *adverse event surveillance*, *decision support alerts* and *automatic summaries*. Creating annotation models is costly. Using fine-grained models for several purposes might be an efficient approach. Our aim is to study whether an existing corpus with fine-grained certainty level annotations can be used for creating multiple scenario-specific certainty level groups, and to study whether limitations in the existing corpus are transferred as limitations in the chosen scenarios. We build automatic classifiers for each scenario, and analyze the results quantitatively. Annotation discrepancies in the corpus are scrutinized and analyzed qualitatively. To our knowledge, no previous research has used fine-grained certainty level annotations for building several use cases with coarse-grained certainty level groups, nor has this been performed on Swedish clinical text.

2 Method

A Swedish clinical corpus annotated for fine-grained certainty levels on a diagnostic statement level was used[1]. The fine-grained classification was collapsed into groups for three different coarse-grained e-health scenarios. Automatic classifiers for each scenario were built, using Conditional Random Fields and simple

[1] Approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm) permission number 2009/1742-31/5.

local context features. Results were evaluated quantitatively through precision, recall and F-score. Annotation discrepancies were analyzed qualitatively through manual corpus analysis.

2.1 Corpus Characteristics

The corpus consists of assessment entries from a medical emergency ward in the Stockholm area. In these entries, reasoning about the patient's status and diseases is documented. Diagnostic statements were automatically tagged in the clinical notes and the annotators judged their certainty levels [9]. An example entry is shown in Figure 1.

> Oklart vad pats symtom kan komma av. Ingen säker <D>infektion</D>. Inga tecken till inflammatorisk sjukdom eller <D>allergi</D>. Reflux med irritation av luftrör och således hosta? Dock har pat ej haft några symtom på <D>refluxesofagit</D>. Ingen ytterligare akut utredning är befogad. Hänvisar till pats husläkare för fortsatt utredning.
> *Unclear what patient's (abbr.) symptoms arise from. No certain <D>infection</D>. No signs of inflammatory disease or <D>allergy</D>. Reflux with irritation of airways and therefore cough? But pat has not had any symptoms of <D>refluxoesophagitis</D>.No further urgent investigation required. Refer to pats GP for continued investigation..*

Fig. 1. Example assessment entry. D = Diagnostic statement. Each marked diagnostic statement was judged for certainty levels. In this case, the diagnostic statements *infektion* (infection), *allergi* (allergy) and *refluxesofagit* (refluxoesophagitis) were to be assigned one of the six certainty level annotation classes.

The annotators were shown the entire assessment entry and were asked to annotate each marked diagnostic statement into one of the six certainty level annotation classes[2]. The certainty levels are modeled in two polarities: *positive* and *negative*, as well as certainty level: *certain*, *probable* or *possible*, see Figure 2. Overall Inter- and Intra Annotator (IAA) results, measured on a subset of the total amount of annotations, were 0.7/0.58 and 0.73/0.6 F-measure/Cohens κ, respectively. This subset was used for the qualitative error analysis. The corpus along with guidelines and further analysis are presented in [9][3].The full corpus consists of 5 473 assessment entries, 6 186 annotated diagnostic statements and 64 832 tokens (7 464 types) annotated by one annotator. Common error types in the annotations are shown in Table 1. We see similarities in both inter- and intra-annotator discrepancies, the most common error type is *1-step* (66% and 69%).

[2] Other classes were also included, but are not analyzed in this work.

[3] The annotators were two senior physicians, accustomed to reading and writing medical records.

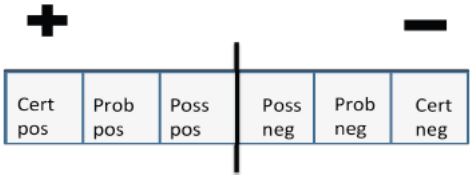

Fig. 2. Fine-grained certainty level classification of diagnostic statements into two polarities and three levels of certainty, in total six classes.

Table 1. The most common error types in the annotated corpus. 1-step = discrepancy in one step, e.g. *certainly negative* vs *probably negative*. Certain/Uncertain = discrepancy between the highest level of certainty and intermediate certainty level classes (*probably* or *possibly*). Polarity = discrepancy in *positive* vs *negative*. n_{inter} = inter-annotator analysis. n_{intra} = intra-annotator analysis

Type	n_{inter}	%	n_{intra}	%
1-step	408	66	284	69
Certain/Uncertain	270	44	191	46
Polarity	99	16	58	14
Total	614	100	411	100

2.2 E-health Scenarios

We define three tentative e-health scenarios: *adverse event surveillance*, *decision support alerts* and *automatic summaries*. These scenarios reflect different needs when it comes to distinguishing and defining the boundaries between certainty levels. The different coarse-grained certainty level groups for the chosen scenarios relate to the original fine-grained classification model as shown in Figure 3. The fine-grained classes *certainly positive*, *probably positive*, *possibly positive*, *possibly negative*, *probably negative* and *certainly negative* are included and excluded in different ways for each scenario. The scenarios are further described below.

Adverse event surveillance One instrument used for surveillance of adverse events in hospital care is the Global Trigger Tool [10]. Here, a number of triggers are defined and used for extraction of records which are subsequently manually scrutinized for adverse events. Automation of the trigger identification procedure and extraction of records saves manual labor, and is presently employed at Karolinska University Hospital for triggers in the structured parts of medical records. Further development of this system would be automatic identification of some of these triggers found in the free-text part of health records, and to this add trigger negation detection. Only cases that are negated with the highest possible level of certainty should be excluded in a potential trigger extraction system. Accurate exclusion of negated cases would lower the overall manual work load. Hence, in this scenario, we get a binary grading: *existence* (at some level of

certainty) or *no existence* (at the most certain level). All five annotation classes except *certainly negative* are collapsed into the *existence* grade.

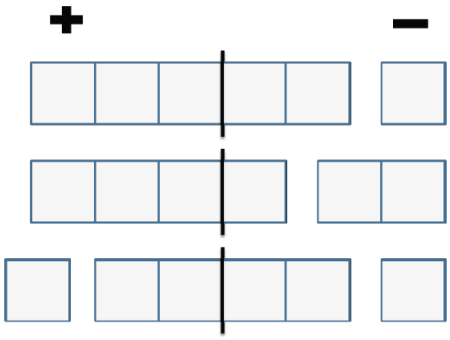

Fig. 3. Modeling e-health use cases by utilizing fine-grained certainty level annotations for coarser-grained classifications, reflecting scenario-specific needs. Top: *adverse event surveillance*. Middle: *decision support alerts*. Bottom: *automatic summaries*.

Decision support alerts In this scenario, the important distinction in an information access setting, is to flag whenever there is a plausible diagnosis [11]. An example of an automated application would be a decision support: if a plausible case is identified, guidelines or other similar recommendations are automatically shown to the clinician in order to take suitable action. Another potential application would be alerting the clinician who is medically responsible for a patient: a nurse documenting a plausible condition produces an automatic alert to the responsible clinician to take action. Separating positive (or near positive) cases from negative cases is important here. Using the fine-grained certainty level annotation classes, we collapse all positive classes as well as *possibly negative*[4] to one group: *plausible existence*. At the negative polarity *probably negative* and *certainly negative* are collapsed into: *no plausible existence*.

Automatic summaries When presented with a new patient, an overview, e.g. textual summary, would help the clinician to get an overall impression of earlier diagnoses and health history. A presentation of diagnoses that have been affirmed, excluded, or discussed as a possibility need to be processed by an automatic information extraction system that can distinguish such cases [12]. Moreover, from a different perspective, patients might be interested in obtaining an overview of their own health records in a similar manner, in order to

[4] The two classes *possibly positive* and *possibly negative* are in this case judged together as a joint middle class.

understand and participate in her or his clinical situation. In this scenario, we use *affirmed* and *negated* as two separate groups, and the remaining intermediate, speculative classes are collapsed into one *speculated* group. Hence, we get a multi-class classification problem with three class labels.

2.3 Automatic Classification and Evaluation

We have used Conditional Random Fields [13], as implemented in CRF++ [5] with default parameter settings for building token level classifiers. All sentences containing diagnostic statements annotated for certainty levels were tokenized[6], and local context features (word, lemma and Part-of-Speech (PoS) tags[7]) with a window of ±4 were used for each token, as this setting produces best results [15]. Each diagnostic statement token was assigned exactly one certainty level class, all other tokens were assigned the class *NONE*.

The corpus was divided into a training set (80%, 4 367 sentences, 4 929 diagnostic statements, 51 523 tokens) and a test set (20%, 1 106 sentences, 1 257 diagnostic statements, 13 309 tokens), with a stratified distribution of annotation class labels, see Table 2.

Table 2. Coarser-grained certainty level annotation class labels, training and test set: number of class instances and percentages in parentheses. S-1 = *adverse event surveillance*. S-2 = *decision support alerts*. S-3 = *automatic summaries*.

		Training set			Test set		
Scenario	Group	S-1 (%)	S-2 (%)	S-3 (%)	S-1 (%)	S-2 (%)	S-3 (%)
S-1	existence	4 372 (89)			1 103 (88)		
	no existence	557 (11)			154 (12)		
S-2	plausible existence		3 934 (80)			995 (80)	
	no plausible existence		995 (20)			262 (20)	
S-3	affirmed			2 463 (50)			625 (50)
	speculated			1 909 (39)			478 (38)
	negated			557 (11)			154 (12)
Total		4 929 (100)	4 929 (100)	4 929 (100)	1 257 (100)	1 257 (100)	1 257 (100)

Results were measured with precision, recall and F-measure, using the CoNLL 2010 Shared task evaluation script conlleval.pl[8]. 95% confidence intervals were calculated for precision and recall. Two baselines were used: majority class baseline and a classifier with no local context features, i.e. the diagnostic statement itself is used as the only feature.

[5] http://crfpp.sourceforge.net/#source

[6] multi-word diagnostic statements such as *heart attack* were concatenated and treated as one token

[7] using a general Swedish tagger [14]

[8] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

3 Results

In this section we present automatic classification results for each e-health scenario, as well as a qualitative error analysis based on the annotated corpus. In the error analysis, we find that difficulties in the distinction between the fine-grained classes *probably negative* and *certainly negative* seem to be the source of most errors in the corpus, and Inter- and Intra-Annotator Agreement (IAA) problems are therefore reflected differently in the three scenarios. We also find that results in the error analysis for the coarse-grained grades are correlated with the distribution of diagnostic statements along the scale of the fine-grained certainty levels. Some diagnostic statements are evenly distributed along this scale, while others are more frequent in the positive polarity (e.g. hypertension, different types of arrythmias, hyperventilation, allergies, different skin diseases) or negative polarity (e.g. thrombosis and ischemia), as shown in [9]. This reflects the clinical need to negate certain disorders in the documentation, but not others. The discrepancies reflect difficulties in judging certainty for different types of diagnostic statements at the respective polarities, with different types of linguistic and clinical assessment problems arising at the respective polarities accordingly.

3.1 Adverse Event Surveillance

In this scenario, we have a binary classification problem: *existence* and *no existence*. This could also be considered similar as a negation detection task.

Classification results In Table 3, results for the baseline (without context features) and for the classifier using a local context window of ±4 is shown. A majority class baseline is 88%. In general, using local context features improves results compared to both baselines (0.89 F-score), but compared to the majority class baseline only a slight improvement is seen. For the minority class *no existence*, context features increase results considerably, in particular for precision (from 0.54 to 0.83), although recall is low (0.51).

Table 3. Classification results for the scenario *adverse event surveillance*. Binary classification: *existence* and *no existence*. P = Precision, R = Recall, F = F-score. 95% confidence intervals are given (±). Majority class baseline = 88%. Baseline = no context features, Local context = word, lemma and PoS-tag, window ±4.

	Baseline			Local context		
Class label	P	R	F	P	R	F
existence	0.53±0.03	0.98±0.01	0.68	0.93±0.01	0.91±0.02	0.92
no existence	0.54±0.08	0.14±0.05	0.23	0.83±0.06	0.51±0.08	0.63
Total	0.53±0.03	0.88±0.02	0.66	0.92±0.01	0.86±0.02	0.89

Error analysis The lower results for *no existence* in the automatic classification for this scenario appears to be connected to known difficulties in the distinction between *probably negative* and *certainly negative* in the annotated corpus. There are not many errors in assigning polarity (see Table 1), i.e. the diagnostic statements are clearly in the negative polarity, but the strength of the negation has been judged differently in many cases. Part of the errors are due to the lexical context surrounding the diagnostic statement. For instance, the phrase *inga hållpunkter för* (no indicators of), has been inconsistently interpreted. These cases are also a source of many errors in the automatic classification. Moreover, these inconsistencies are often related to diagnostic statements belonging to diagnosis types that are difficult to exclude, such as *DVT* (deep venous thrombosis), where complete exclusion is clinically difficult. Speculations arise around these diagnosis types because of important severe consequences if missed or misjudged. There are also inconsistencies that depend on whether the annotator(s) have judged the local or global context (i.e. the whole assessment entry, or only the current sentence). Modifiers such as *liten*, e.g. *liten misstanke* (small suspicion), are an interesting source of errors: these can be interpreted differently depending on whether emphasis is put on *misstanke* (suspicion), or *liten* (small), and would need to be defined further in the guidelines.

3.2 Decision Support Alerts

In this scenario we need two groups. The classification task is hence modeled with binary class labels: *plausible existence* and *no plausible existence*.

Classification results In Table 4, results are shown for the classification baseline as well as for using local context features. A majority class assignment is 80%. Overall results are improved using local context features (from 0.61 F-score to 0.91), and are also improved compared to the majority class baseline. For the minority class *no plausible existence*, results are considerably improved both for precision (from 0.72 to 0.92) and recall (from 0.22 to 0.79).

Table 4. Classification results for the scenario *alerts for decision support*. Binary classification: *plausible existence* and *no plausible existence*. P = Precision, R = Recall, F = F-score. 95% confidence intervals are given (±). Majority class baseline = 80%. Baseline = no context features, Local context = word, lemma and PoS-tag, window ±4.

	Baseline			Local context		
Class label	P	R	F	P	R	F
plausible existence	0.48±0.03	0.97±0.01	0.64	0.95±0.01	0.90±0.02	0.92
no plausible existence	0.72±0.05	0.22±0.05	0.34	0.92±0.03	0.79±0.05	0.85
Total	0.49±0.03	0.82±0.02	0.61	0.94±0.01	0.88±0.02	0.91

Error analysis The boundary in the fine-grained classification model is shifted towards the positive polarity, as compared to the *adverse event surveillance* scenario. The main source of errors lies in cases where certain clinical exclusion is very difficult, due to the nature of the diagnosis itself (e.g. *DVT*). Another source of errors lies in cases where tests have been performed in order to exclude a specific diagnosis. These cases are difficult since performing a test in itself is an indication that there is a risk of this diagnosis, but from the surrounding context it can be evident that the diagnosis is highly unlikely.

3.3 Automatic Summaries

In this scenario, we need three grades, resulting in a multi-class classification problem: *affirmed*, *speculated*, and *negated*.

Classification results A majority class assignment (*affirmed*) is 50%. In Table 5 results for the classifiers (baseline, and context window ±4) are shown. Using local context features result in a considerable improvement for all classes (0.8 F-score, overall average, compared to 0.5, both baselines). Recall for *negated* is, however, relatively low (0.55).

Table 5. Classification results for the scenario *automatic summary*. Multi-class classification: *affirmed*, *speculated* and *negated*. P = Precision, R = Recall, F = F-score. 95% confidence intervals are given (±). Majority class baseline = 50%. Baseline = no context features, Local context = word, lemma and PoS-tag, window ±4.

	Baseline			Local context		
Class label	P	R	F	P	R	F
affirmed	0.79±0.03	0.72±0.03	0.75	0.87±0.03	0.81±0.03	0.84
speculated	0.25±0.02	0.77±0.02	0.38	0.81±0.02	0.77±0.02	0.79
negated	0.50±0.08	0.18±0.08	0.27	0.81±0.06	0.55±0.08	0.66
Total	0.40±0.03	0.67±0.03	0.50	0.84±0.02	0.76±0.02	0.80

Error analysis In this scenario, we focus on an error analysis in the positive polarity, which is not covered in the other two scenarios. These errors mostly reflect difficulties in distinguishing between *probably positive* and *certainly positive* in the annotated corpus. A majority of the cases are due to linguistic markers such as *misstänkt* <*D*>*x*</*D*> (suspected <D>x</D>) or *kliniska tecken på* <*D*>*x*</*D*> (clinical signs of <D>x</D>). We see more discrepancies in the annotations concerning diagnosis types determined by subjective judgement, e.g. *hyperventilering* (hyperventilation) and *panikångest* (panic disorder) than diagnosis types that are measured objectively, e.g. *hypertoni* (hypertension). A difference in the judgments made by the human annotators lies in whether they

have based their judgments on clinical knowledge or linguistic markers, e.g. *Urinprov pos. därför troligen urinvägsinf.* (Urine sample pos. thus probably urinary tract inf.) We observe some difficult cases for chronic diseases. For instance, the example *troligen stressutlöst astma* (probably stress triggered asthma), could be interpreted as *certainly positive* in the sense that the patient is diagnosed with asthma, or as *probably positive* in the sense that this particular event of an asthma attack is probably triggered by stress.

4 Analysis and Discussion

In this study we present work using a corpus annotated with fine-grained certainty classes on a diagnostic statement level, for coarser-grained e-health scenarios. We present three scenarios: *adverse event surveillance*, *decision support alerts* and *automatic summaries*. These scenarios are real-world situations where computerized support is beneficial [12], and where Natural Language Processing techniques involving negation handling may be useful [11]. Each scenario requires different certainty level models, and we collapse classes from the fine-grained classification model into three different coarser-grained groups. We build classifiers using local context features for each scenario. A qualitative analysis on annotation errors deepens the understanding of problems in the boundaries between certainty level classes. We observe promising results by the automatic classifiers for all three scenarios (0.89 F-score (*adverse event surveillance*), 0.91 F-score (*decision support alerts*) and 0.8 F-score (*summaries*), overall average). Our main findings are that it is feasible to use a fine-grained certainty level classification model of diagnostic statements for building coarser-grained e-health scenarios. Although overall IAA is relatively low for the fine-grained model [9], most errors are found in the 1-step borders between the fine-grained levels, thus yielding higher IAA for coarser-grained situations. Annotation discrepancies in intermediate certainty level classes do not pose problems when classes are collapsed into coarser-grained certainty level groups. However, there are some problematic issues, in particular in the distinction between *probably negative* and *certainly negative* in the fine-grained classification model, which need to be further defined in the annotation guidelines. This problem becomes evident when looking at the results for the automatic classifier for the scenario *adverse event surveillance*, where recall in the minority class *no existence* is 0.51. Whether the fine-grained model is considered a sliding scale, or a two-step decision (polarity followed by certainty level) by the annotators is also a factor that should be studied further and need to be clarified when creating fine-grained certainty level annotation tasks.

Previous work (e.g. [1], [2], [4], [5]), on similar tasks are difficult to compare for several reasons. For instance, the certainty level models, annotation tasks, corpora and classification approaches are different to those employed in this work. However, some general trends are observed, such as the problem of skewed class distributions and ambiguity of context cues. Interestingly, local context features in a window of ± 4 are shown to be useful also for English [1], as well as

for Swedish [15]. Cross-lingual studies would be a very interesting continuation of this work. Moreover, the fine-grained certainty levels might also be useful as features for other (higher-level) classification tasks.

Qualitative studies on terminologies used for expressing diagnostic certainties reveal that intermediate probabilities are more often difficult to agree on among human (clinical) evaluators ([16] and [17]), which is in line with our observations. This is an inherently subjective task, and it is not trivial to define what upper performance bounds would be for classifiers.

4.1 Limitations

The automatic classifiers have been built on annotations by one annotator only, not on a consensus set by several annotators. Overall results are also affected by skewed class distributions, results for minority classes need to be further analyzed. Moreover, other classification algorithms should be tested. We treat this task as a token level classification problem, using Conditional Random Fields for classification. Other classification algorithms or representations might be better suited for this task, this should be studied further and compared. More detailed feature analysis is also needed, as well as under- or oversampling data for dealing with the problem of skewed class distributions. For instance, no global context features have been used, nor any clinical domain-knowledge based features, such as test results.

Moreover, the qualitative error analysis is performed on annotations by two annotators, and only on a subset of the original corpus. A correlation between inter-annotator discrepancies and the errors resulting from the classifiers should be analyzed in future studies.

4.2 Significance of Study

Our results are valuable for further work on creating accurate information extraction methods for clinical real-world cases. In health care, there is a constant need for quick decisions based on earlier documentation. This is often complicated by the accumulating mass of text surrounding every patient case. Automatic text processing for applications such as decision support and summaries or overviews, adapted to natural language, would facilitate the clinical workday. Also, automation of surveillance tools for adverse events can assist in improvement of hospital care. This study indicates that it is possible to use a general resource for specific scenario solutions. Instead of creating, in this case, three coarse-grained annotation tasks and subsequent corpora, one fine-grained model can be used for several purposes successfully. To our knowledge, no previous research has used fine-grained certainty level annotations for building several coarse-grained use cases, nor has this been studied on Swedish clinical text.

Acknowledgments We would like to express our appreciations to the anonymous and known reviewers for invaluable comments and suggestions for this paper.

References

1. Uzuner, Ö., Zhang, X., Sibanda, T.: Machine Learning and Rule-based Approaches to Assertion Classification. JAMIA **16** (2009) 109–115
2. Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W.: ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics **42** (2009) 839–851
3. 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** (2008)
4. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. JAMIA **18** (2011) 552–556
5. de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J., Zhu, X.: Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. JAMIA **18** (2011) 557–562
6. Wilbur, J.W., Rzhetsky, A., Shatkay, H.: New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics **7** (2006) 356+
7. Rubin, V.L., Liddy, E.D., Kando, N.: Certainty identification in texts: Categorization model and manual tagging results. In: Computing Affect and Attitutde in Text: Theory and Applications. Springer (2006)
8. Saurí, R.: A Factuality Profiler for Eventualities in Text. PhD thesis, Brandeis University (2008)
9. Velupillai, S., Dalianis, H., Kvist, M.: Factuality Levels of Diagnoses in Swedish Clinical Text. In Moen, A., Andersen, S.K., Aarts, J., Hurlen, P., eds.: Proc. XXIII Intl. Conf. of the European Federation for Medical Informatics, Oslo, IOS Press (2011) 559 – 563
10. F.A., G., Resar, R.: IHI Global Trigger Tool for Measuring Adverse Events (Second Edition). IHI Innovation Series white paper. Cambridge, Massachusetts: Institute for Healthcare Improvement (2009)
11. Denny, J.C., Miller, R.A., Waitman, L.R., Arrieta, M.A., Peterson, J.F.: Identifying QT prolongation from ECG impressions using a general-purpose Natural Language Processor. IJMI **78 S 1** (2009) S34–S42
12. Kvist, M., Skeppstedt, M., Velupillai, S., Dalianis, H.: Modeling human comprehension of Swedish medical records for intelligent access and summarization systems, a physician's perspective. In: Proc. 9th Scandinavian Conf. on Health Informatics, SHI, Oslo (2011)
13. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML. (2001) 282–289
14. Knutsson, O., Bigert, J., Kann, V.: A robust shallow parser for Swedish. In: Proceedings of Nodalida 2003, Reykavik, Iceland (2003)
15. Velupillai, S.: Automatic Classification of Factuality Levels – A Case Study on Swedish Diagnoses and the Impact of Local Context. In: Proc. 4th Intl. Symp. on Languages in Biology and Medicine (LBM 2011), Singapore (2011)
16. Khorasani, R., Bates, D.W., Teeger, S., Rotschild, J.M., Adams, D.F., Seltzer, S.E.: Is terminology used effectively to convey diagnostic certainty in radiology reports? Academic Radiology **10** (2003) 685–688
17. Hobby, J.L., Tom, B.D.M., Todd, C., Bearcroft, P.W.P., Dixon, A.K.: Communication of doubt and certainty in radiological reports. The British Journal of Radiology **73** (2000) 999–1001