Automatic Classification of Factuality Levels – A Case Study on Swedish Diagnoses and the Impact of Local Context

Sumithra Velupillai
Dept. of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100, SE-16440 Kista, Sweden
sumithra@dsv.su.se

Abstract

Clinicians express different levels of knowledge certainty when reasoning about a patient's status. Automatic extraction of relevant information is crucial in the clinical setting, which means that factuality levels need to be distinguished. We present an automatic classifier using Conditional Random Fields, which is trained and tested on a Swedish clinical corpus annotated for factuality levels at a diagnosis statement level: the Stockholm EPR Diagnosis-Factuality Corpus. The classifier obtains promising results (best overall results are 0.699 average F-measure using all classes, 0.762 F-measure using merged classes), using simple local context features. Preceding context is more useful than posterior, although best results are obtained using a window size of ± 4. Lower levels of certainty are more problematic than higher levels, which was also the case for the human annotators in creating the corpus. A manual error analysis shows that conjunctions and other higher-level features are common sources of errors.

1 Introduction and Background

Decision-making is a central task in clinical work, which involves complex reasoning based on information at hand. Clinicians are faced with new patients and need to be able to assess the patient's status according to several criteria, depending on situation, clinical expertise, previous history, patient descriptions, etc. Clinicians document their findings and reasoning in words, either through dictation or directly in written form. Today, most documentation is inserted in digitized systems, where information is stored both in structured and unstructured (free text) forms. One of the central activities in clinical work is the process of diagnosing. A clinician needs to classify what (possible) problem(s) a patient suffers from. This process involves much reasoning. Since clinicians document a large amount in free text, there is a lot of information to be extracted that could be of use in the decision-making process, e.g. similar cases and overviews. For this, accurate information extraction techniques are needed.

In many situations, it is not clear what disease a patient actually suffers from. A physician might receive insufficient background information, symptoms might be unclear or there might be several alternative possibilities to a patient's status. Moreover, it may also be the case that a disease is excluded as a possibility. Such reasoning is documented in free text, and these distinctions are crucial to model if an information extraction system is to be built for retrieving diagnostic information from clinical documentation.

The importance of modeling modality and negation for information extraction and information access purposes has been recognized in several different research areas lately, e.g. in the biomedical domain, for opinion mining and subjectivity analysis, summarization, text mining. Different models for representing modality and negation have been proposed, ranging from analyzing sentence levels to event levels, exploiting specific surface markers (keywords and phrases) or more complex linguistic constructions. When it comes to building automatic systems for distin-

guishing factuality levels, we see two general approaches: rule-based or machine learning models exploiting annotated corpora.

Annotation models: Wilbur et al. (2006) present a model of five qualitative dimensions for characterizing scientific articles: focus, polarity, certainty, evidence and directionality. The aim is to be able to identify reliable scientific facts, or informative fragments, along these dimensions. This model is applied on a sentence level (or subsentential if the sentence is complex). Polarities are modeled on a positive and negative axis, and certainty levels are modeled on a scale of 0 – 3, where 0 indicates complete uncertainty. The highest degree (3) represents complete certainty. Similarly, Rubin et al. (2006) create an annotation scheme where degree (certainty level), perspective (whose certainty), focus (object of certainty) and time is modeled. Certainty levels are modeled on four levels: absolute, high, moderate and low. Here, polarity is not included in the model. This model is applied on newspaper articles.

A different approach is presented in FactBank (Saurí, 2008), where factuality levels are annotated on an event level. Moreover, factuality is modeled on two different polarities: *positive* and *negative*, followed by certainty levels *certainly*, *probably* and *possibly*. Linguistically motivated markers are discussed in detail. For cases where polarity cannot be ascertained, *underspecified* is used. This corpus consists of newspaper articles, as a second layer on top of TimeBank (Pustejovsky et al., 2006).

Automatic systems: The BioScope corpus (Vincze et al., 2008) is a manually annotated corpus containing biomedical texts as well as clinical free-text (radiology reports), annotated for negation and speculation cues (token level) along with their linguistic scope (sentence level). This corpus has been used for the development of supervised learning classifiers, and was used in the CoNLL 2010 Shared task (Farkas et al., 2010), where the top performing system obtained an F-measure of 0.864 for detecting uncertain sentences (Tang et al., 2010), and 0.573 for detecting in-sentence hedge cues (Morante et al., 2010).

In the clinical domain, rule-based systems for distinguishing negations and uncertainties have been successfully developed, e.g. Harkema et al. (2009) and Friedman et al. (2004). ConText (Harkema et al., 2009) is an extension of the NegEx algorithm (Chapman et al., 2001), where negated, historical, hypothetical conditions, and conditions not experienced by the patient are automatically identified in emergency department reports. RadReportMiner (Wu et al., 2009) is a context-aware search engine, taking into account negations and uncertainties, achieving improved precision results (0.81) compared to a generic search engine (0.27) using a modified version of the NegEx algorithm, including expanded sets of negation and uncertainty keywords.

Studies on uncertainty expressions in the clinical domain: Verbal and numerical uncertainty expressions and their role in communicating clinical information have been studied from many perspectives and for different purposes, e.g. decision-making, interpretation, impact on physicians, patients and information systems. Most often, studies have used direct and indirect scaling procedures, where a fixed number of verbal expressions are given for judgment, and evaluating results by inter- and intra-subject agreement (see e.g. Clark (1990) for a critical review). In general, intra-evaluator agreement is found to be high, and inter-evaluator agreement to be low. Intermediate probabilities are often more difficult to agree on, while very high or low probabilities result in higher agreement (see Khorasani et al. (2003), Hobby et al. (2000), Christopher and Hotz (2004)). In many cases, the main conclusion is to recommend the use of controlled vocabularies for expressing different levels of certainty. The verbal expressions range from one word expressions such as *definite*, *likely*, *possible* to longer expressions such as *cannot be excluded*. The relationship between expressing probabilities verbally or numerically has also been studied (e.g. Timmermans (1994) and Renooij and Witteman (1999)), where findings suggest that verbal expressions are found to be more vague than numerical, and hence more difficult to use in decision-making.

The work presented here is divided into the following parts: 1) automatically classifying factuality levels using the Stockholm EPR Diagnosis-Factuality Corpus (Velupillai et al., 2011) with local context features and 2) evaluating by measur-

ing precision, recall and F-measure and 3) performing a qualitative, manual error analysis. To our knowledge, no previous research have modeled factuality levels in clinical assessment documentation on a diagnostic statement level, nor on Swedish clinical documentation.

2 Methods

Our work process is: (1) automatic classification of the Stockholm EPR Diagnosis-Factuality Corpus using local context features and (2) evaluating the classification results quantitatively (precision, recall and F-measure) and qualitatively by manual error analysis[1].

2.1 The Stockholm EPR Diagnosis-Factuality corpus

> Låg sannolikhet för <D>dvt</D> pga frånvaro av riskfaktorer och blygsamma klin. fynd.
> *Low probability for <D>dvt</D> (abbr.) due to lack of risk factors and modest clinical (abbr.) findings..*

Example 1: Example sentence from the Stockholm EPR Diagnosis-Factuality Corpus, D = Diagnostic statement. In this case, the diagnostic statement *dvt* (deep venous thrombosis) was to be annotated for factuality level, e.g. *possibly positive*.

The Stockholm EPR Diagnosis-Factuality corpus consists of documents that have been extracted from a university hospital emergency ward included in the Stockholm EPR Corpus (Dalianis et al., 2009). The documents are extracted from a medical emergency ward, since this is a type of clinic where several different types of diseases can be encountered. Only entries documented under the category *Bedömning* (Assessment) have been used. This entry type was chosen since it is where most reasoning, speculation and discussion regarding the patient status is documented. Each assessment entry is saved as one document, i.e. no other information from the patient record is used in the annotation task. Two domain experts (A1 and A2); senior physicians, both accustomed to reading and writing Swedish medical records, annotated the diagnostic statements for factuality levels. The largest set was annotated by A1, which is used in the presented work. Inter- and Intra-Annotator Agreement (IAA) results are 0.7/0.58 F-measure and 0.73/0.6 Cohens κ, respectively. The corpus is further described in (Velupillai et al., 2011).

[1]This research has been carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2009/1742-31/5.

	Total (#)	With diagnoses (#)
Documents	3 846	
Sentences	26 232	5 741
Tokens	283 007	69 355
Types (lemmas)	14 834	6 077
Diagnoses	6 483	
Diagnosis types (lemmas)	302	

Table 1: General statistics: Stockholm EPR Diagnosis-Factuality Corpus. Total set annotated by annotator A1. Each assessment entry is one document. Each diagnostic statement is one annotation instance. Punctuation is included in tokens and types.

2.1.1 Corpus characteristics

In the Stockholm EPR Diagnosis-Factuality Corpus, sentence and keyword level annotations are not used. Instead, only diagnostic statements are annotated for factuality levels. A manually created list of diagnostic statements was used, including different inflections, variants and abbreviations. The diagnostic statements in this list were automatically marked in brackets for the annotators to assign factuality levels. The whole assessment entry was shown to the annotators. An example sentence is shown in Example 1. General statistics of the Stockholm EPR Diagnosis-Factuality Corpus are shown in Table 1.

Following the factuality modeling presented in (Saurí, 2008), factuality levels are first defined in two polarities: *Positive* and *Negative*. Each of these were also graded: *Certain*, *Probable* or *Possible*. In total six annotation classes are used for marking factuality levels. Furthermore, the annotation class *Not Diagnosis* is used for cases where, e.g. the diagnostic statement in its context was something else (e.g. infektion (*infection*, short for clinical department)), or kol (*coal* in its meaning medical coal, not the diagnosis *COPD*). The annotation class *Other* is also included for cases where e.g. the diagnostic statement referred to someone other than the patient, or where the annotator could not assess the diagnostic statement

according to any of the other classes[2]. The resulting annotation classes were the result of thorough discussions between the annotators and the research group. Guidelines for the annotation task are publicly available[3].

2.1.2 Class distributions

The distribution of factuality level annotation classes is shown in Table 2. *Certainly positive* is in clear majority, almost 50%. *Possibly negative*, *Not diagnosis* and *Other* are very rare, with less than 3%, respectively. The negative polarity amounts to 21.7% in total, and intermediate positive factuality levels (probably and possibly) amount to 26.2%, which means that a fair amount of diagnostic statements are speculative or negated. Thus, distinguishing factuality levels is very important for accurate information extraction.

Annotation Class	*n*	%
Certainly Positive	3 088	47.6
Probably Positive	1 039	16.0
Possibly Positive	663	10.2
Possibly Negative	139	2.2
Probably Negative	546	8.4
Certainly Negative	711	11.0
Not Diagnosis	117	1.8
Other	180	2.8
Σ	6 483	100.0

Table 2: Class distributions.

As a broad coverage approach was chosen, several different diagnostic statements are present in the annotated set. In Table 3, we see example distributions per class for some of the most frequent diagnostic statements. We observe that some diagnostic statements are more commonly used only in one class, e.g. förmaksflimmer (*atrial fibrillation*) and hypertoni (*hypertension*): *certainly positive* (93% and 89%, respectively). On the other hand, dvt (*deep venous thrombosis*, abbreviated), and infektion (*infection*) are more spread out and can be discussed in all factuality levels and polarities. Infektion (*infection*) is also sometimes used for mentioning a clinic, which is why it can be annotated as *not diagnosis*. Ischemi (*ischaemia*) is almost always assigned a negative polarity annotation class (28% *probably negative*, 58% *certainly negative*).

2.2 Automatic classification

For automatic classification we have used Conditional Random Fields (CRF) (Lafferty et al., 2001), as implemented in CRF++ [4], a classification algorithm that has been successful for similar Natural Language Processing (NLP) classification tasks. We use default settings, with no added parameter tuning. As there are cases where there are several diagnostic statements in one sentence, we do not treat this as a sentence level classification task. Instead, each token in all sentences containing an annotation instance (the assigned factuality level class for the marked diagnostic statement[5]) is classified. We have, in this work, used local features surrounding each annotation instance.

Many previous studies on expressions of probabilities in the clinical domain have used specific keywords and phrases within a small context window (e.g. Khorasani et al. (2003), Hobby et al. (2000)). Although these studies have been used in English settings, we found similar patterns in our Swedish clinical corpus. We limit the context window to ± 4. For expanding the language model, we also use Part-of-Speech (PoS) tags and lemmas, extracted from a general language tagger for Swedish (Knutsson et al., 2003). We use simple features: word, lemma and PoS tag.

All results from the automatic classification experiments were measured on a test set containing 20% of the total amount of annotations. 80% of the total set is used for training. Approximately the same proportions of annotation class distributions are used in both sets. Results were measured with precision, recall and F-measure, using the CoNLL 2010 Shared task evaluation script conlleval.pl[6]. 95% confidence intervals were calculated for precision and recall.

[2] This class can be considered as a *neutral* class, for cases where no polarity and factuality level can be assessed (underspecified in Saurí (2008)).

[3] http://www.dsv.su.se/hexanord/guidelines/guidelines_stockholm_epr_diagnosis_factuality_corpus.pdf

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

[5] diagnostic statements that are multiword tokens, such as *angina pektoris* are concatenated into one token.

[6] http://www.cnts.ua.ac.be/conll2000/chunking/output.html

Diagnosis	CP	PrP	PoP	PoN	PrN	CN	ND	O
deep venous thrombosis								
dvt	83 (21)	36 (9)	89 (23)	25 (6)	91 (23)	55 (14)	0	12 (3)
infection								
infektion	74 (25)	41 (14)	40 (13)	8 (3)	49 (16)	55 (18)	18 (6)	13 (4)
atrial fibrillation								
förmaksflimmer	241 (93)	6 (2)	5 (2)	0	0	3 (1)	0	5 (2)
hypertension								
hypertoni	213 (89)	16 (7)	5 (2)	0	0	0	0	4 (2)
ischaemia								
ischemi	2 (2)	1 (1)	11 (9)	2 (2)	32 (28)	67 (58)	0	1 (1)

Table 3: Example distributions diagnostic statements vs factuality level classes, *n* (%). CP = Certainly Positive, PrP = Probably Positive, PoP = Possibly Positive, PoN = Possibly Negative, PrN = Probably Negative, CN = Certainly Negative, ND = Not Diagnosis and O = Other

3 Results

The training set consists of 4 583 sentences, 5 171 annotation instances, and the test set of 1 158 sentences, 1 312 annotation instances. In these initial experiments, we are interested in looking at the local context, which is why we use only those sentences that contain annotated instances. For evaluating the automatic classification results, we use as a baseline the word itself as the only feature. Following the IAA-results (see Velupillai et al. (2011)), where the intermediate factuality levels often were a source of lower results, we also perform automatic classification where we merge the two intermediate factuality level classes per polarity, i.e. *probably/possibly positive/negative* are merged into *probably possibly positive* and *probably possibly negative*. We also merge *other* and *not diagnosis* into one class in order to increase the number of instances. Majority class baseline is also used for evaluating results.

All instances that are not annotated are assigned the class *NONE*. Baseline results are shown in Table 4. The majority class *certainly positive* obtains relatively high results (0.742 (all classes) and 0.758 (merged classes) F-measure). Overall average results for all classes is 0.561 F-measure and 0.605 for merged classes, an improvement over the majority class baseline (47.6% for all classes as well as merged classes).

3.1 Local context features

Using the closest context (window $\pm$ 1) improves results considerably compared to the baseline (using only the word itself) for all classes and settings (0.659 F-measure, all annotation classes, 0.704 F-measure, merged classes), using only words and lemmas. Intermediate classes in the positive polarity gain from merging, while *not diagnosis* obtains lower results. Increasing the window size step by step improves results further, and best results are obtained using a window size of $\pm$ 4, with words, lemmas and PoS information (Table 5). Using only words, lemmas and PoS information in a four-span window *preceding* the word itself yields similar results (0.69 F-measure for all classes, 0.736 for merged classes), indicating that preceding context is extremely valuable. Contrasting with posterior features ($\pm$4) yields lower results: 0.599 (all classes) and 0.649 (merged classes). PoS information is useful in combination with words and/or lemmas, not as a feature on its own. A considerable improvement is seen when increasing the window size from $\pm$2 to $\pm$3 (0.67 to 0.69 all classes, 0.716 to 0.737 merged classes). The greatest gain is seen for *certainly negative*, with an increase in F-measure from 0.546 to 0.674 (all classes) and 0.55 to 0.676.

3.2 Error analysis

The erroneous classification results from using window $\pm$ 4 with CRF classification have been analyzed (semi-)manually. The most frequent errors are misclassifications within the same polarity, or missed instances. We observed some general trends:

- Conjunctions: in many cases, conjunctions such as *och* (and), *eller* (or), cause errors, indicating that surface level features are problematic; these instances might have been captured if syntactic information was used, e.g. *Inga hållpunkter i lab och ekg för pågående* **ischemi** (No basis in lab and ecg for ongoing **ischaemia**).

	P_a (95% CI)	R_a (95% CI)	F_a	P_m (95% CI)	R_m (95% CI)	F_m	Merged
CP	0.657 ± 0.04	0.853 ± 0.03	0.742	0.736 ± 0.03	0.781 ± 0.03	0.758	CP
PrP	0.5 ± 0.07	0.3202 ± 0.06	0.390	0.478 ± 0.05	0.611 ± 0.05	0.536	PrPoP
PoP	0.464 ± 0.08	0.09 ± 0.05	0.151				
PoN	0.0 ± 0.0	0.0 ± 0.0	0.0	0.377 ± 0.08	0.153 ± 0.06	0.217	PrPoN
PrN	0.273 ± 0.08	0.296 ± 0.09	0.284				
CN	0.393 ± 0.08	0.487 ± 0.08	0.435	0.454 ± 0.08	0.351 ± 0.07	0.396	CN
O	0.0 ± 0.0	0.0 ± 0.0	0.0	1.0 ± 0.0	0.2545 ± 0.11	0.40	O-ND
ND	1.0 ± 0.0	0.433 ± 0.18	0.605				
Avg_a	0.565 ± 0.03	0.557 ± 0.03	0.561	0.609 ± 0.03	0.601 ± 0.03	0.605	Avg_m

Table 4: CRF++, Baseline, i.e. only using the word itself as feature, without surrounding context. $_a$ = all annotation classes: CP = Certainly Positive, PrP = Probably Positive, PoP = Possibly Positive, PoN = Possibly Negative, PrN = Probably Negative, CN = Certainly Negative, ND = Not Diagnosis and O = Other. $_m$ = merged classes: PrPoP = Probably and Possibly positive, PrPoN = Probably and Possibly Negative, O-ND = Other and Not diagnosis. P = Precision, R = Recall, F = F-measure, CI = Confidence Interval.

- Lab results: in some cases, lab results (or similar) seem to be highly indicative for specific diagnoses, but are not frequent enough in the training set or captured well in this model.
- Short sentences: in some cases, the sentence only contained the diagnostic statement itself, where the reasoning was documented in the remaining document. Here, it is evident that larger contexts may be very important.
- Longer discussions: for some diagnostic statements, a long discussion preceded the diagnostic statement itself, with many modifiers and speculations. In these cases, the local window did not model the factuality level well.

4 Discussion

In this study we present experiments on the impact of local features for an automatic factuality level classifier of Swedish diagnostic statements using the Stockholm EPR Diagnosis-Factuality Corpus. Using local context features improves results, in particular for annotation classes in the positive polarity, as well as for *certainly negative*. Preceding features are very valuable, both on their own and in combination with posterior features. Posterior features are not useful on their own. PoS information in combination with words and/or lemmas contributes to slight improvements. More complex language models are probably needed for improving results in the infrequent classes where context plays a larger role, as shown in the error analysis. Using syntactic features such as dependency parses and rules for linguistic constructions might be useful here (see e.g. Kilicoglu and Bergler (2008) and Velldal et al. (2010)). Moreover, in some cases we observe the need for including features at a cross-sentence level, and the inclusion of other types of features such as laboratory results. Some phrases might reflect ambiguous uses; for some diagnostic statements they are used for indicating high levels of certainty while for other diagnostic statement types they are used for indicating speculation. This is worth investigating further.

Merging annotation classes is fruitful for obtaining improved results, especially in the positive polarity. The same trends are not evident for the negative polarity, which might be due to the fact that the number of instances is much lower. Moreover, the two *possibly* levels were often confused even for the same annotator. These very low certainty levels might instead be merged into one *neutral* or *very low certainty* class, where polarity is not as important. The classes *not diagnosis* and *other* are too different to merge. Successful classification of the annotation class *other* probably needs more sophisticated language modeling, such as co-reference resolution, in the cases where instances are diagnostic statements referring to someone other than the patient.

In this corpus, we have a large amount of different diagnostic statement types. Grouping these and classifying factuality levels according to diagnostic statement type might lead to the insight that different types of features are indicative for different types of diagnostic statements. Moreover, the different annotation classes might also benefit from class-specific feature modeling, as was seen for *certainly negative*, where using the preceding context as features gave the best results.

	P_a (95% CI)	R_a (95% CI)	F_a	P_m (95% CI)	R_m (95% CI)	F_m	Merged
CP	**0.826** $\pm$ 0.03	**0.814** $\pm$ 0.03	0.82	**0.839** $\pm$ 0.03	**0.818** $\pm$ 0.03	0.828	CP
PrP	0.64 $\pm$ 0.07	0.576 $\pm$ 0.07	0.604	**0.825** $\pm$ 0.04	0.72 $\pm$ 0.05	0.769	PrPoP
PoP	0.643 $\pm$ 0.08	0.437 $\pm$ 0.08	0.521				
PoN	0.636 $\pm$ 0.20	0.304 $\pm$ 0.18	0.412	0.58 $\pm$ 0.08	0.55 $\pm$ 0.08	0.564	PrPoN
PrN	0.504 $\pm$ 0.09	0.528 $\pm$ 0.09	0.516				
CN	**0.789** $\pm$ 0.06	0.584 $\pm$ 0.08	**0.716**	0.79 $\pm$ 0.06	0.604 $\pm$ 0.08	0.686	CN
O	0.444 $\pm$ 0.19	0.16 $\pm$ 0.14	0.25				
ND	1.0 $\pm$ 0.0	0.6 $\pm$ 0.18	0.75	0.885 $\pm$ 0.08	0.418 $\pm$ 0.13	0.568	O-ND
Avg	0.744 $\pm$ 0.02	0.66 $\pm$ 0.03	0.699	0.805 $\pm$ 0.02	0.723 $\pm$ 0.02	0.762	Avg_m

Table 5: CRF++, window $\pm$ 4, word, lemma and PoS. $_a$ = all annotation classes: CP = Certainly Positive, PrP = Probably Positive, PoP = Possibly Positive, PoN = Possibly Negative, PrN = Probably Negative, CN = Certainly Negative, ND = Not Diagnosis and O = Other. $_m$ = merged classes: PrPoP = Probably and Possibly positive, PrPoN = Probably and Possibly Negative, O-ND = Other and Not diagnosis. P = Precision, R = Recall, F = F-measure, CI = Confidence Interval.

4.1 Limitations

The study design has some limitations. The concept of a diagnostic statement is not trivial and given the limited collection of diagnostic statements created in this work, the distribution of diagnostic statements might not reflect a real-world scenario. However, with this corpus, we have material for analyzing differences and similarities in how different diseases and diagnosis types are expressed with regards to factuality levels. We have shown that there are patterns among diagnostic statements, these should be analyzed further.

A further limitation of this model and the resulting corpus is the low number of annotations in some annotation classes. Merging intermediate probability levels improved results in the positive polarity, but in the negative polarity the same trend could not be observed. Here, we also had a much lower amount of instances. *Possibly negative* was a difficult class even for the same annotator, and might need further definitions. Moreover, the annotation class *other* is very complex, as it can be used for diagnostic statements referring to someone other than the patient. For these types of instances, co-reference resolution is needed, and adding further levels to the annotation model such as *perspective* or *source* might be useful (see e.g. Saurí (2008), Wilbur et al. (2006) and Rubin et al. (2006)). As the overall IAA results are relatively low Velupillai et al. (2011), further refinements in guidelines and resolving conflicting annotations to build a consensus corpus would be useful.

There are also limitations in the classification design; we have not tuned any parameters, nor have we compared with other learning algorithms. This should be further studied. Moreover, in order to increase the number of annotations and extending the corpus, active learning techniques could be very useful. The factuality level model with in total six levels of certainty could be considered as a continuum or scale, not necessarily as mutually independent classes. From this point of view, the factuality classification might be modeled differently, for instance through treating factuality as a continuous variable.

4.2 Significance of study

To our knowledge, no other studies have approached the study of factuality levels on a diagnosis basis in clinical Swedish. Our results show that the created model is feasible for an annotation task, resulting in a corpus that can be used for automatic classification. We see that speculative expressions in Swedish clinical assessments to a large extent are fairly consistent within a small context window, but that for improving results further, deeper language and feature models and might be needed. Automatic factuality level classification could be integrated in an information extraction system for clinicians and clinical researchers, where different factuality levels are distinguished. Choosing a broad approach gives further knowledge in how similarities and differences between different factuality levels among diagnostic statements in Swedish are expressed.

Acknowledgments

The author wishes to thank the two domain expert annotators Mia Kvist, M. D., PhD, and Prof. Gunnar Nilsson, M. D. for their valuable work on the creation of the annotation guidelines and the annotated corpus.

References

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. 2001. Evaluation of negation phrases in narrative clinical reports. In *AMIA Symp*, pages 105–109.

M. M. Christopher and C. S. Hotz. 2004. Cytologic diagnosis: expression of probability by clinical pathologists. *Veterinary Clinical Pathology*, 33(2):84–95.

D. A. Clark. 1990. Verbal Uncertainty Expressions: A Critical Review of Two Decades of Research. *Current Psychology: Research & Reviews*, 9(3):203–235.

H. Dalianis, M. Hassel, and S. Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proc. ISHIMR 2009*, Kalmar, Sweden, October 14-16. Awarded best paper.

R. Farkas, V. Vincze, G. Móra, J. Csirik, and G. Szarvas. 2010. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proc. 14th CoNLL*, pages 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. 2004. Automated Encoding of Clinical Documents Based on Natural Language Processing. *Journal of the American Medical Informatics Association*, 11(5):392–402.

H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomedical Informatics*, 42:839–851.

J. L. Hobby, B. D. M. Tom, C. Todd, P. W. P. Bearcroft, and A. K. Dixon. 2000. Communication of doubt and certainty in radiological reports. *The British Journal of Radiology*, 73:999–1001, September.

R. Khorasani, D. W. Bates, S. Teeger, J. M. Rotschild, D. F. Adams, and S. E. Seltzer. 2003. Is terminology used effectively to convey diagnostic certainty in radiology reports? *Academic Radiology*, 10:685–688.

H. Kilicoglu and S. Bergler. 2008. Recognizing Speculative Language in Biomedical Research Articles: A Linguistically Motivated Perspective. In *Proc. BioNLP 2008*, pages 38–45, June.

O. Knutsson, J. Bigert, and V. Kann. 2003. A robust shallow parser for Swedish. In *Proc. Nodalida 2003*, Reykavik, Iceland.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289.

R. Morante, V. Van Asch, and W. Daelemans. 2010. Memory-based resolution of in-sentence scopes of hedge cues. In *Proc. 14th CoNLL*, pages 40–47, Uppsala, Sweden, July.

J. Pustejovsky, M. Verhagen, R. Saurí, J. Littman, R. Gaizauskas, G. Katz, I. Mani, R. Knippen, and A. Setzer. 2006. Timebank 1.2. Linguistic Data Consortium (LDC). Philadelphia, PA.

S. Renooij and C. Witteman. 1999. Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, 22:169–194.

V. L. Rubin, E. D. Liddy, and N. Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing Affect and Attitutde in Text: Theory and Applications*. Springer.

R. Saurí. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University.

B. Tang, X. Wang, X. Wang, B. Yuan, and S. Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proc. 14th CoNLL*, pages 13–17, Uppsala, Sweden, July.

D. Timmermans. 1994. The Roles of Experience and Domain of Expertise in Using Numerical and Verbal Probability Terms in Medical Decisions. *Medical Decision Making*, 14:146–156.

E. Velldal, L. Øvrelid, and S. Oepen. 2010. Resolving speculation: Maxent cue classification and dependency-based scope rules. In *Proc. 14th CoNLL*, pages 48–55, Uppsala, Sweden, July.

S Velupillai, H. Dalianis, and M. Kvist. 2011. Factuality levels of diagnoses in swedish clinical text. In *Proceedings of MIE 2011*, pages 559–563. IOS Press, August.

V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S-11).

J. W. Wilbur, A. Rzhetsky, and H. Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356+, July.

A. S. Wu, B. H. Do, J. Kim, and D. L. Rubin. 2009. Evaluation of negation and uncertainty detection and its impact on precision and recall in search. *J Digit Imaging*.