How Certain are Clinical Assessments? Annotating Swedish Clinical Text for (Un)certainties, Speculations and Negations

Hercules Dalianis, Sumithra Velupillai
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100, 164 40 Kista, Sweden
E-mail: {hercules, sumithra}@dsv.su.se

Abstract

Clinical texts contain a large amount of information. Some of this information is embedded in contexts where e.g. a patient status is reasoned about, which may lead to a considerable amount of statements that indicate uncertainty and speculation. We believe that distinguishing such instances from factual statements will be very beneficial for automatic information extraction.
We have annotated a subset of the Stockholm Electronic Patient Record Corpus for certain and uncertain expressions as well as speculative and negation keywords, with the purpose of creating a resource for the development of automatic detection of speculative language in Swedish clinical text. We have analyzed the results from the initial annotation trial by means of pairwise Inter-Annotator Agreement (IAA) measured with F-score. Our main findings are that IAA results for certain expressions and negations are very high, but for uncertain expressions and speculative keywords results are less encouraging. These instances need to be defined in more detail. With this annotation trial, we have created an important resource that can be used to further analyze the properties of speculative language in Swedish clinical text. Our intention is to release this subset to other research groups in the future after removing identifiable information.

1. Introduction

The use of electronic patient records (EPRs) is increasing in the healthcare sector, which leads to a growing amount of digitalized data. Automatic methods for accessing information from such data is an important research area.
The Stockholm Electronic Patient Corpus (Stockholm EPR Corpus) is a clinical corpus containing over one million patient records, encompassing 2 000 clinics from the Stockholm area stretching over the years 2006 to 2008 (Dalianis et al., 2009). The Stockholm EPR Corpus contains both structured information and unstructured information (free text). The free text entries are semi-structured, since the free text is entered under several free text categories, for example *Bedömning (Assessment), Aktuell status (Current status), Social Bakgrund (Social Background)*.
In EPRs, the patient status is described and reasoned about. We believe that this leads to a considerable amount of statements that indicate uncertainty and speculation, where clinicians describe situations that are difficult to confirm. Distinguishing such instances from factual, or certain, instances is important if the information is to be extracted automatically, since the former alters the meaning of the expression. In the long run, systems for Information Extraction, Information Retrieval or Knowledge Discovery may be improved by including such distinctions, where, for instance, a clinician would benefit from accessing information about previous, similar cases when faced with a difficult situation.
We have annotated a subset of the Stockholm EPR corpus for certain and uncertain expressions as well as speculative and negation keywords, with the purpose of creating a resource for the development of automatic detection of speculative language in Swedish clinical text.
Our aim is to analyze the results from the initial annotation trial by means of pairwise Inter-Annotator Agreement (IAA) measured with F-score. Our intention is to release this subset to other research groups in the future, after ensuring that no identifiable information is included in the subset.

2. Previous research

Research on the identification of speculative language, or "hedging", has gained a large amount of interest lately, especially for scientific articles and abstracts in the biomedical domain. Research findings often contain tentative results, where further analysis might be needed. Distinguishing such findings from factual statements is crucial for information extraction systems. Several research groups have analyzed the characteristics of speculative language in biomedical scientific writings.
Light et al. (2004) found 11 percent speculative language in Medline abstracts from scientific articles in Biomedicine. Here, four annotators annotated 891 sentences each as either highly speculative, low speculative, or definite. Their Inter-Annotator Agreement (IAA) results, measured with kappa, ranged between 0.54 and 0.68. They also found that the majority of the speculative sentences appeared towards the end of the abstract. Finally they also annotated a larger set of sentences (the last two sentences in all annotated data sets, (i.e. the last two sentences in the abstracts)) containing in total 2 093 sentences and found 18 percent speculative sentences and 82 percent definite sentences.
In the BioScope corpus (Vincze et al., 2008), both medical (clinical) free texts, biological full papers and biological scientific abstracts have been annotated,

This research has been carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2007/1625-31/5.

encompassing more than 20 000 sentences, where over 10 percent of the sentences were either speculative or negated. We are specifically interested in the results for the clinical sub-corpus where 13.55 percent of the clinical texts contained negation and 13.99 percent contained speculative keywords However, the authors do not report any results of whether negation and speculation keywords co-occur. In Kilicoglu & Bergler (2008), non-lexical features for identifying speculative language are used (as well as lexical cues). Some of these are defined as negated non-speculative ("unhedging") cues, such as *no evident*. They report promising results on the automatic identification of speculative language in biomedical research articles.

The IAA results in the clinical sub-corpus of BioScope for negation keywords ranged between 91 and 96 percent F-score, and for speculative keywords the results ranged between 84 and 92 percent F-score. These results indicate that negation keywords seem to be easier to identify than speculation keywords. In Light et al. (2004), methods for automatic identification of speculative language by using annotated corpora have also been developed. Using Support Vector Machines (SVM), and evaluating with 10 fold cross evaluation, 84 percent precision and 39 percent recall was obtained.

Özgür & Radev (2009) used two parts of the annotated BioScope corpus, namely the biological full papers and biological scientific abstracts (9 full text papers and 1 273 abstracts) for automatic identification of speculative language. They also used SVM for two classification tasks: identifying keywords used in speculative context, and determining the scope of these keywords. For scientific abstracts they obtained 88.16 percent recall, 95.21 percent precision and 91.50 F-score. They also found that speculative keywords co-occur and that they are more common in the *Conclusion* and *Discussion* parts of the articles.

Morante & Daelemans (2009) describe work on the same two classification tasks on all three BioScope subcorpora. Here, different machine learning methods are used for the different tasks, including SVM, Memory-based learning and Conditional Random Fields (CRF). Overall, the results for abstracts and papers are considerably higher than for clinical text for the first classification task, which influences results on the second classification task. These results show that differences in text type are important to consider.

3. Method

We have annotated 6 740 randomly extracted sentences from the Stockholm EPR corpus, from the free text category *Bedömning* (Assessment). Three annotators with no prior knowledge of the content worked on the task; one senior level student (SLS), one undergraduate computer scientist (UCS), and one undergraduate language consultant (ULC).

In order to make the corpus comparable, we developed guidelines similar to those for the BioScope corpus (Vincze et al., 2008). However, in the BioScope corpus, certain expressions, as well as expressions containing question marks (*?*), were not annotated. The following annotation classes were used in the work presented here: *Certain_expression, Uncertain_expression, Negation, Speculative_words, Undefined_expression* and *Undefined_speculative_words*.

For each randomly extracted sentence, the full free text entry was shown to the annotators, in order for them to see the context of the sentence. (See Examples 1 and 2 below). Sentences were extracted using a simple tokenizing strategy based on regular expressions. Each sentence had to be judged either as a certain, uncertain or undefined expression. In cases where a sentence contained both, for instance through subordinate clauses, a sentence could be broken into sub-expressions. Within these expressions, negated or speculative keywords were annotated if present. By doing this, both sentence level and token level annotations were captured. We did not, however, in this annotation trial, include annotations for the scope of a token level speculative or negated keyword, i.e. those syntactic units that are modified by the keyword In even intervals (in total seven), during the three working weeks, the group of annotators met to discuss the task. This was carried out in order to measure IAA results over time and after resolving problems, similar to Haverinen et al. (2009).

4. Results

We have measured IAA by pair wise F-score, treating one set of annotations as the gold standard for each combination of annotator pairs. As a final result, we give the average result. We have measured both partial and exact matching. Exact matching is at token level while partial matching is at character level. In Tables 1, 2, 3, 4 and 5, results are shown.

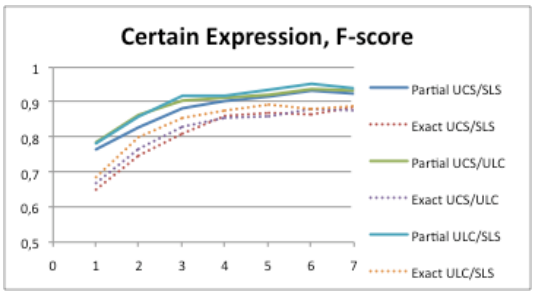

Table 1. Results for Certain Expression, pairwise IAA (F-score) over time both for partial and exact matching.

Looking at sentence level, the IAA results for *Certain_expression* were in general very high (0.84 F-score for exact matches) while considerably lower for *Uncertain_expression*, see Tables 1 and 2.

Having discussions among the annotators in time intervals yields an improvement for results on

Certain_expression, we also see a convergence between partial and exact matching, see Table 1. For *Uncertain_expression*, results over time are very disparate. However, we also see a tendency for convergence between partial and exact matching, specifically between annotation intervals 5 to 7, which is probably due to the discussions among the annotators, see Table 2.

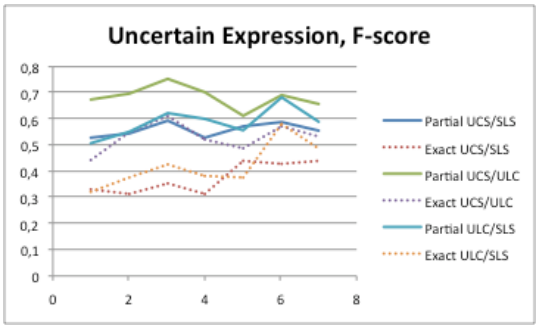

Table 2. Results for Uncertain Expression, pairwise IAA (F-score) over time both for partial and exact matching.

The annotation class *Speculative_words* obtains low IAA results in general, see Table 3. However, we also see a tendency for convergence between partial and exact matching and discrepancies between exact and partial matches are lower here, which shows that larger scopes for annotating uncertain expressions, see Table 2, are more difficult to define.

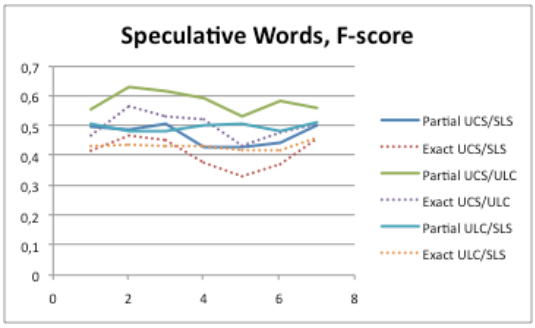

Table 3. Results for Speculative Words, pairwise IAA (F-score) over time both for partial and exact matching.

Negation keywords obtained very high IAA results (over 0.80 F-score), for these there was also an improvement in results over time. The total average results, see Table 4, show an overall improvement over time.

These results are, however, heavily influenced by the dominance of the annotation class *Certain_expression*.

Looking at the actual contents of the annotations, from the 6 740 sentences, we find an average total amount of 6 996 annotated expressions. On average, 13.5 percent of these are annotated as uncertain expressions (ranging between 11.8 and 15.7 percent). The average amount of annotated speculative words was 1 624 and the average amount of negation keywords was 1 008.

Looking at the token level annotations *Speculative_words* and *Negations*, the average amount of unique keywords was 538 and 13, respectively. The most common speculative keywords for all three annotators were unigrams such as *sannolikt* (*likely*) and *möjligen* (*possibly*).

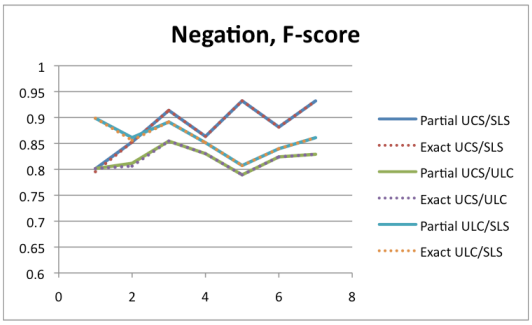

Table 4. Results for Negation Words, pairwise IAA (F-score) over time both for partial and exact matching.

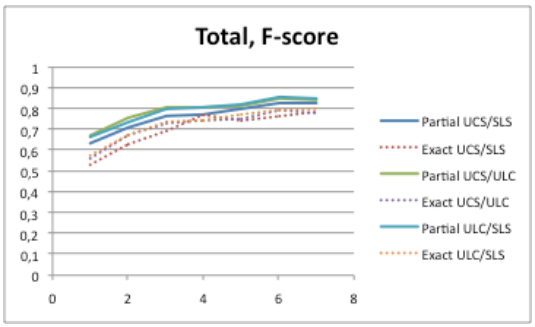

Table 5. Total average results, pairwise IAA (F-score) over time both for partial and exact matching.

However, 52 percent (on average) of the speculative keywords were unigrams, the rest being *n*-grams of varying length. Moreover, several annotations of speculative keywords included negations, such as *ingen misstanke* (*no suspicion*) and *inga tydliga tecken* (*no clear signs*). Many of these conform well to those listed as indicative features of speculative language in Kilicoglu & Bergler (2008), where negated "unhedgers" form speculative cues. Notable is also that the negation keywords only included evident negation words such as *inte* (*not*) and *inga* (*none*). In Swedish, it is also possible to negate words with prefixes such as *o-*, as in *oklar* (*indistinct*). No such words were annotated as negation keywords by the annotators.

When looking at the sentences and the contexts in which they were annotated, there was a great variety in how large the context was, how long the sentences

were, and in what setting they were written. In Example 1 we see a sentence that is annotated as an *uncertain expression*. This sentence contains a negation combined with a non-speculative keyword, which together form a multi-word annotation (*Speculative_words*), making the whole expression uncertain. In this example, we also see that the simple sentence tokenization created an annotation instance that had to be broken down into two sub-expressions.

Bedömning:
<sentence_1>
<Uncertain_expression>Statusmässigt
<Speculative_words><Negation>inga
</Negation> säkra</Speculative_words>
artriter</Uncertain_expression>.
<Certain_expression>Lungrtg Huddinge ua
</Certain_expression>.</sentence>
Leverprover ua.

Translation to English:

Assessment:
<sentence_1>
<Uncertain_expression>Status-wise
Speculative_words><Negation>no
</Negation> certain</Speculative_words>
arthritis</Uncertain_expression>.
<Certain_expression>cxr Huddinge woco
</Certain_expression>.</sentence>
Liver samples woco

Example 1. An annotated sentence containing a negation and a certain expression making the whole expression uncertain.

Example 2 shows an annotated sentence within a context that contains a relatively large amount of reasoning, concerning several issues regarding the patient status, giving a more thorough account of the level of certainty (please cf. with the reasoning processes in Groopman (2007)).

Bedömning:
<sentence_2><Uncertain_expression>
Har lite <Speculative_words> undringar
</Speculative_words> om brakyterapi
<Speculative_words> kunde vara
</Speculative_words> aktuellt i hans fall
</Uncertain_expression>.</sentence> Har
haft den diskussionen uppe med Bengt
Karlsson. Jag har svårt och tro det eftersom
han går på Onkologen och rimligtvis hade man
tänkt på den behandlingen om man hade ansett
att det finns möjlighet men jag lovar att
skriva ett brev till Lars Olof Svensson om
detta. Vad det gäller pricken på mandibeln
verkar det mest som ett lite aterom tycker
jag men det är klart att hudmetastas är ju
inte uteslutet. Jag lämnar dock den frågan
helt till Onkologen.

Translation to English:

Assessment:
<sentence_2><Uncertain_expression>
I have some <Speculative_words> concerns
</Speculative_words> about whether
brachytherapy <Speculative_words> could be
</Speculative_words> considered in his case
</Uncertain_expression>.</sentence> I have
had that discussion with Bengt Karlsson. I
have difficulties believing this since he is
treated at the Oncology clinic and they must
have considered this treatment if they
thought this was possible, but I promise to
write a letter to Lars Olof Svensson
regarding this. Regarding the mark on the
mandible, I think it mostly seems to be a bit
of aterom but of course a Cutaneous
metastasis can not be excluded. However, I
leave that question entirely to the Oncology
clinic.

Example 2. An annotated sentence within a context that contains a relatively large amount of reasoning.

Regarding sensitive information that can reveal the identity of a patient, the annotators identified in total 15 personal names (from the total amount of 290 085 tokens). One half consisted of personal names of clinical personnel, and the other half consisted of patient first personal names. Moreover, seven social security numbers were found. This indicates that personal names are extremely rare in the Assessment field (0.02 per thousand). In the Stockholm EPR PHI Corpus (another subset of the Stockholm EPR Corpus), consisting of 380 000 tokens (containing all the free text entry fields), 0.19 per thousand patient first personal names were found. However, in this corpus, no social security numbers were identified (Dalianis & Velupillai 2010). Although identifiable information seems to be very infrequent, it is crucial to ensure that no identifiable information about an individual is kept if a corpus is to be released for further research.

5. Discussion

We have presented initial results on an annotation trial for speculative language in Swedish clinical texts. Our main findings are that IAA results for certain expressions and negations are very high, but for uncertain expressions and speculative words results are less encouraging. These instances need to be defined in more detail.

Our results are comparable to those presented in Light et al. (2004). However, our annotations of certainties and negations obtain high IAA results and the training effect is significant. Our IAA results are lower than Vincze et al. (2008), but this may be due to differences in corpora. In the clinical sub-corpus presented in Vincze et al. (2008), radiology reports are annotated, while the annotations presented here were randomly extracted from all clinics in the Stockholm EPR corpus. Moreover, the sentences extracted for this annotation

trial were extracted from the free-text entries under the heading *Bedömning* (*Assessment*). This heading may be used differently in different health care units, and may hence contain diverse types of statements. The sentences could, for instance, contain descriptions of the current, overall status of a patient or a short-term plan for medication. It was also evident that expressions of speculations may differ greatly between clinical disciplines.

During discussions among the annotators, some specific properties were pointed out. One was the question of perspective; the patient's and the physician's, especially for uncertain expressions. After annotation interval 2 it was decided to only annotate the physician's perspective. Another point was the level of (un)certainty; many expressions were more or less (un)certain. A grading of four scales was proposed: *Completely certain*, *Quite certain*, *Quite uncertain* and *Completely uncertain*. Such a distinction would probably have a great effect on the sentences currently annotated as *Certain_expression*, which, in the current set, in the majority of cases, merely indicate that the sentence is *not* uncertain. Furthermore, vagueness was often difficult to distinguish from uncertainty.

5. Conclusions and future work

The research presented here is to our knowledge the first work carried out on annotating speculations in clinical text written in Swedish. It is also the first time that both certain and uncertain expressions have been explicitly annotated.

Although IAA results for speculative words and uncertain expressions were low, we believe that the identification of such language is important for future Information Access research. However, further definitions are needed. In particular, the distinction between different perspectives in uncertain expressions is important and needs to be handled. This distinction is probably a specific property of EPRs and probably not present to the same extent in scientific text.

Moreover, looking at different health care disciplines, there may be great differences in how uncertainties and speculations are expressed. This is particularly interesting when looking at specific diagnoses, e.g. speculations about certain diagnoses such as brain tumors are probably very rare, while speculations about for instance psychiatric diagnoses may be much more common. We will analyze the annotated set by dividing it into different health care units, in order to analyze whether such differences are apparent.

We plan to analyze the annotations further, by looking in more detail at the speculative words, investigating their characteristics, analyzing the multi-word expressions, finding out to what extent they are combined with negations and what implications this has, as well as analyzing in which part of the text the uncertain expressions are present. When it comes to negation keywords, we plan to analyze them in particular for finding which constructions where they, combined with non-speculative keywords, form a speculative expression. We will also analyze the scopes of the annotated keywords, in order to identify what expressions they modify. Moreover, we plan to create a consensus corpus from the annotated set presented here, to use for training and testing a machine learning system on our annotations, to investigate the possibilities of automatic classification. For such a system, we will look at syntactic patterns as well as word-level features.

6. Acknowledgements

We would like to thank our three annotators, Helen Allvin, Freja Dalianis and Aron Henriksson for their endurance and accuracy during their work and also for the interesting and fruitful discussions during the annotation sessions

7. References

Dalianis, H., M. Hassel and S. Velupillai. 2009. The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of, the 14th International Symposium for Health Information Management Research (ISHIMR 2009),* Kalmar, Sweden, 14-16 October, 2009, pp 243-249. Awarded Best Paper.

Dalianis, H. and S. Velupillai. 2010. De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields. To be published in Journal of Biomedical Semantics..

Groopman, J. 2007. *How doctors think*, Houghton Mifflin Company, New York.

Kilicoglu, H. and S. Bergler 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(S-11).

Haverinen, K., F. Ginter, V. Laippala and T. Salakoski. 2009. Parsing Clinical Finnish: Experiments with Rule-Based and Statistical Dependency Parsers. In *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*.

Light, M., X. Y. Qiu, and P. Srinivasan. 2004. The Language of Bioscience: Facts, Speculations, and Statements in Between. In *BioLINK 2004: Linking Biological Literature, Association for Computational Linguistics Ontologies, and Databases*, pp. 17-24.

Morante, R. and Daelemans, W. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on BioNLP*, pp 28-36, Boulder, Colorado, June 2009.

Vincze V., Szarvas G, Farkas R, Móra G, and Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes *BMC Bioinformatics. 2008; 9 (Suppl 11): S9*. Published online 2008 November 19. doi: 10.1186/1471-2105-9-S11-S9.