Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus

Hercules Dalianis, Maria Skeppstedt
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100
SE-164 40 Kista, Sweden
{hercules, mariask}@dsv.su.se

Abstract

In this paper we describe the creation of a consensus corpus that was obtained through combining three individual annotations of the same clinical corpus in Swedish. We used a few basic rules that were executed automatically to create the consensus. The corpus contains negation words, speculative words, uncertain expressions and certain expressions. We evaluated the consensus using it for negation and speculation cue detection. We used Stanford NER, which is based on the machine learning algorithm Conditional Random Fields for the training and detection. For comparison we also used the clinical part of the BioScope Corpus and trained it with Stanford NER. For our clinical consensus corpus in Swedish we obtained a precision of 87.9 percent and a recall of 91.7 percent for negation cues, and for English with the Bioscope Corpus we obtained a precision of 97.6 percent and a recall of 96.7 percent for negation cues.

1 Introduction

How we use language to express our thoughts, and how we interpret the language of others, varies between different speakers of a language. This is true for various aspects of a language, and also for the topic of this article; negations and speculations. The differences in interpretation are of course most relevant when a text is used for communication, but it also applies to the task of annotation. When the same text is annotated by more than one annotator, given that the annotating task is non-trivial, the resulting annotated texts will not be identical. This will be the result of differences in how the text is interpreted, but also of differences in how the instructions for annotation are interpreted. In order to use the annotated texts, it must first be decided if the interpretations by the different annotators are similar enough for the purpose of the text, and if so, it must be decided how to handle the non-identical annotations.

In the study described in this article, we have used a Swedish clinical corpus that was annotated for certainty and uncertainty, as well as for negation and speculation cues by three Swedish-speaking annotators. The article describes an evaluation of a consensus annotation obtained through a few basic rules for combining the three different annotations into one annotated text.[1]

2 Related research

2.1 Previous studies on detection of negation and speculation in clinical text

Clinical text often contains reasoning, and thereby many uncertain or negated expressions. When, for example, searching for patients with a specific symptom in a clinical text, it is thus important to be able to detect if a statement about this symptom is negated, certain or uncertain.

The first approach to identifying negations in Swedish clinical text was carried out by Skeppstedt (2010), by whom the well-known NegEx algorithm (Chapman et al., 2001), created for English clinical text, was adapted to Swedish clinical text. Skeppstedt obtained a precision of 70 percent and a recall of 81 percent in identifying negated diseases and symptoms in Swedish clinical text. The NegEx algorithm is purely rule-based, using lists of cue words indicating that a preceding or following disease or symptom is negated. The English version of NegEx (Chapman et al., 2001) obtained a precision of 84.5 percent and a recall of 82.0 percent.

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

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 5–13,
Uppsala, July 2010.

Another example of negation detection in English is the approach used by Huang and Lowe (2007). They used both parse trees and regular expressions for detecting negated expressions in radiology reports. Their approach could detect negated expressions both close to, and also at some distance from, the actual negation cue (or what they call negation signal). They obtained a precision of 98.6 percent and a recall of 92.6 percent.

Elkin et al. (2005) used the terms in SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms), (SNOMED-CT, 2010) and matched them to 14 792 concepts in 41 health records. Of these concepts, 1 823 were identified as negated by humans. The authors used Mayo Vocabulary Server Parsing Engine and lists of cue words triggering negation as well as words indicating the scope of these negation cues. This approach gave a precision of 91.2 percent and a recall of 97.2 percent in detecting negated SNOMED-CT concepts.

In Rokach et al. (2008), they used clinical narrative reports containing 1 766 instances annotated for negation. The authors tried several machine learning algorithms for detecting negated findings and diseases, including hidden markov models, conditional random fields and decision trees. The best results were obtained with cascaded decision trees, with nodes consisting of regular expressions for negation patterns. The regular expressions were automatically learnt, using the LCS (longest common subsequence) algorithm on the training data. The cascaded decision trees, built with LCS, gave a precision of 94.4 percent, a recall of 97.4 percent and an F-score of 95.9 percent.

Szarvas (2008) describes a trial to automatically identify speculative sentences in radiology reports, using Maximum Entropy Models. Advanced feature selection mechanisms were used to automatically extract cue words for speculation from an initial seed set of cues. This, combined with manual selection of the best extracted candidates for cue words, as well as with outer dictionaries of cue words, yielded an F-score of 82.1 percent for detecting speculations in radiology reports. An evaluation was also made on scientific texts, and it could be concluded that cue words for detecting speculation were domain-specific.

Morante and Daelemans (2009) describe a machine learning system detecting the scope of negations, which is based on meta-learning and is trained and tested on the annotated BioScope Corpus. In the clinical part of the corpus, the authors obtained a precision of 100 percent, a recall of 97.5 percent and finally an F-score of 98.8 percent on detection of cue words for negation. The authors used TiMBL (Tilburg Memory Based Learner), which based its decision on features such as the words annotated as negation cues and the two words surrounding them, as well as the part of speech and word forms of these words. For detection of the negation scope, the task was to decide whether a word in a sentence containing a negation cue was either the word starting or ending a negation scope, or neither of these two. Three different classifiers were used: support vector machines, conditional random fields and TiMBL. Features that were used included the word and the two words preceding and following it, the part of speech of these words and the distance to the negation cue. A fourth classifier, also based on conditional random fields, used the output of the other three classifiers, among other features, for the final decision. The result was a precision of 86.3 percent and a recall of 82.1 percent for clinical text. It could also be concluded that the system was portable to other domains, but with a lower result.

2.2 The BioScope Corpus

Annotated clinical corpora in English for negation and speculation are described in Vincze et al. (2008), where clinical radiology reports (a subset of the so called BioScope Corpus) encompassing 6 383 sentences were annotated for negation, speculation and scope. Henceforth, when referring to the BioScope Corpus, we only refer to the clinical subset of the BioScope Corpus. The authors found 877 negation cues and 1 189 speculation cues, (or what we call speculative cues) in the corpora in 1 561 sentences. This means that fully 24 percent of the sentences contained some annotation for negation or uncertainty. However, of the original 6 383 sentences, 14 percent contained negations and 13 percent contained speculations. Hence some sentences contained both negations and speculations. The corpus was annotated by two students and their work was led by a chief annotator. The students were not allowed to discuss their annotations with each other, except at regular meetings, but they were allowed to discuss

with the chief annotator. In the cases where the two student annotators agreed on the annotation, that annotation was chosen for the final corpus. In the cases where they did not agree, an annotation made by the chief annotator was chosen.

2.3 The Stanford NER based on CRF

The Stanford Named Entity Recognizer (NER) is based on the machine learning algorithm Conditional Random Fields (Finkel et al., 2005) and has been used extensively for identifying named entities in news text. For example in the CoNLL-2003, where the topic was language-independent named entity recognition, Stanford NER CRF was used both on English and German news text for training and evaluation. Where the best results for English with Stanford NER CRF gave a precision of 86.1 percent, a recall of 86.5 percent and F-score of 86.3 percent, for German the best results had a precision of 80.4 percent, a recall of 65.0 percent and an F-score of 71.9 percent, (Klein et al., 2003). We have used the Stanford NER CRF for training and evaluation of our consensus.

2.4 The annotated Swedish clinical corpus for negation and speculation

A process to create an annotated clinical corpus for negation and speculation is described in Dalianis and Velupillai (2010). A total of 6 740 randomly extracted sentences from a very large clinical corpus in Swedish were annotated by three non-clinical annotators. The sentences were extracted from the text field Assessment (*Bedömning* in Swedish). Each sentence and its context from the text field Assessment were presented to the annotators who could use five different annotation classes to annotate the corpora. The annotators had discussions every two days on the previous days' work led by the experiment leader.

As described in Velupillai (2010), the annotation guidelines were inspired by the BioScope Corpus guidelines. There were, however, some differences, such as the scope of a negation or of an uncertainty not being annotated. It was instead annotated if a sentence or clause was certain, uncertain or undefined. The annotators could thus choose to annotate the entire sentence as belonging to one of these three classes, or to break up the sentence into subclauses.

Pairwise inter-annotator agreement was also measured in the article by Dalianis and Velupillai (2010) . The average inter-annotator agreement increased after the first annotation rounds, but it was lower than the agreement between the annotators of the BioScope Corpus.

The annotation classes used were thus *negation* and *speculative words*, but also *certain expression* and *uncertain expression* as well as *undefined*. The annotated subset contains a total of 6 740 sentences or 71 454 tokens, including its context.

3 Method for constructing the consensus

We constructed a consensus annotation out of the three different annotations of the same clinical corpus that is described in Dalianis and Velupillai (2010). The consensus was constructed with the general idea of choosing, as far as possible, an annotation for which there existed an identical annotation performed by at least two of the annotators, and thus to find a majority annotation. In the cases where no majority was found, other methods were used.

Other options would be to let the annotators discuss the sentences that were not identically annotated, or to use the method of the BioScope Corpus, where the sentences that were not identically annotated were resolved by a chief annotator (Vincze et al., 2008). A third solution, which might, however, lead to a very biased corpus, would be to not include the sentences for which there was not a unanimous annotation in the resulting consensus corpus.

3.1 The creation of a consensus

The annotation classes that were used for annotation can be divided into two levels. The first level consisted of the annotation classes for classifying the type of sentence or clause. This level thus included the annotation classes *uncertain*, *certain* and *undefined*. The second level consisted of the annotation classes for annotating cue words for negation and speculation, thus the annotation classes *negation* and *speculative words*. The annotation classes on the first level were considered as more important for the consensus, since if there was no agreement on the kind of expression, it could perhaps be said to be less important which cue phrases these expressions contained. In the following constructed example, the annotation tag *Uncertain* is thus an annotation on the first level, while the annotation tags *Negation* and *Speculative words* are on the second level.

```
<Sentence>
    <Uncertain>
      <Speculative_words>
        <Negation>Not</Negation>
        really
      </Speculative_words>
      much worse than before
    </Uncertain>
 <Sentence>
```

When constructing the consensus corpus, the annotated sentences from the first rounds of annotation were considered as sentences annotated before the annotators had fully learnt to apply the guidelines. The first 1 099 of the annotated sentences, which also had a lower inter-annotator agreement, were therefore not included when constructing the consensus. Thereby, 5 641 sentences were left to compare.

The annotations were compared on a sentence level, where the three versions of each sentence were compared. First, sentences for which there existed an identical annotation performed by at least two of the annotators were chosen. This was the case for 5 097 sentences, thus 90 percent of the sentences.

For the remaining 544 sentences, only annotation classes on the first level were compared for a majority. For the 345 sentences where a majority was found on the first level, a majority on the second level was found for 298 sentences when the scope of these tags was disregarded. The annotation with the longest scope was then chosen. For the remaining 47 sentences, the annotation with the largest number of annotated instances on the second level was chosen.

The 199 sentences that were still not resolved were then once again compared on the first level, this time disregarding the scope. Thereby, 77 sentences were resolved. The annotation with the longest scopes on the first-level annotations was chosen.

The remaining 122 sentences were removed from the consensus. Thus, of the 5 641 sentences, 2 percent could not be resolved with these basic rules. In the resulting corpus, 92 percent of the sentences were identically annotated by at least two persons.

3.2 Differences between the consensus and the individual annotations

Aspects of how the consensus annotation differed from the individual annotations were measured. The number of occurrences of each annotation class was counted, and thereafter normalised on the number of sentences, since the consensus annotation contained fewer sentences than the original, individual annotations.

The results in Table 1 show that there are fewer uncertain expressions in the consensus annotation than in the average of the individual annotations. The reason for this could be that if the annotation is not completeley free of randomness, the class with a higher probability will be more frequent in a majority consensus, than in the individual annotations. In the cases where the annotators are unsure of how to classify a sentence, it is not unlikely that the sentence has a higher probability of being classified as belonging to the majority class, that is, the class *certain*.

The class *undefined* is also less common in the consensus annotation, and the same reasoning holds true for *undefined* as for *uncertain*, perhaps to an even greater extent, since *undefined* is even less common.

Also the *speculative* words are fewer in the consensus. Most likely, this follows from the *uncertain* sentences being less common.

The words annotated as *negations*, on the other hand, are more common in the consensus annotation than in the individual annotations. This could be partly explained by the choice of the 47 sentences with an annotation that contained the largest number of annotated instances on the second level, and it is an indication that the consensus contains some annotations for negation cues which have only been annotated by one person.

Type of Annot. class	**Individ.**	**Consens.**
Negation	853	910
Speculative words	1 174	1 077
Uncertain expression	697	582
Certain expression	4 787	4 938
Undefined expression	257	146

Table 1: Comparison of the number of occurrences of each annotation class for the individual annotations and the consensus annotation. The figures for the individual annotations are the mean of the three annotators, normalised on the number of sentences in the consensus.

Table 2 shows how often the annotators have divided the sentences into clauses and annotated each clause with a separate annotation class. From the table we can see that annotator A and also an-

notator H broke up sentences into more than one type of the expressions *Certain*, *Uncertain* or *Undefined expressions* more often than annotator F. Thereby, the resulting consensus annotation has a lower frequency of sentences that contained these annotations than the average of the individual annotations. Many of the more granular annotations that break up sentences into certain and uncertain clauses are thus not included in the consensus annotation. There are instead more annotations that classify the entire sentence as either *Certain*, *Uncertain* or *Undefined*.

Annotators	**A**	**F**	**H**	**Cons.**
No. sentences	349	70	224	147

Table 2: Number of sentences that contained more than one instance of either one of the annotation classes *Certain*, *Uncertain* or *Undefined expressions* or a combination of these three annotation classes.

3.3 Discussion of the method

The constructed consensus annotation is thus different from the individual annotations, and it could at least in some sense be said to be better, since 92 percent of the sentences have been identically annotated by at least two persons. However, since for example some expressions of uncertainty, which do not have to be incorrect, have been removed, it can also be said that some information containing possible interpretations of the text, has also been lost.

The applied heuristics are in most cases specific to this annotated corpus. The method is, however, described in order to exemplify the more general idea to use a majority decision for selecting the correct annotations. What is tested when using the majority method described in this article for deciding which annotation is correct, is the idea that a possible alternative to a high annotator agreement would be to ask many annotators to judge what they consider to be certain or uncertain. This could perhaps be based on a very simplified idea of language, that the use and interpretation of language is nothing more than a majority decision by the speakers of that language.

A similar approach is used in Steidl et al. (2005), where they study emotion in speech. Since there are no objective criteria for deciding with what emotion something is said, they use manual classification by five labelers, and a majority voting for deciding which emotion label to use. If less than three labelers agreed on the classification, it was omitted from the corpus.

It could be argued that this is also true for uncertainty, that if there is no possibility to ask the author of the text, there are no objective criteria for deciding the level of certainty in the text. It is always dependent on how it is perceived by the reader, and therefore a majority method is suitable. Even if the majority approach can be used for subjective classifications, it has some problems. For example, to increase validity more annotators are needed, which complicates the process of annotation. Also, the same phenomenon that was observed when constructing the consensus would probably also arise, that a very infrequent class such as *uncertain*, would be less frequent in the majority consensus than in the individual annotations. Finally, there would probably be many cases where there is no clear majority for either completely certain or uncertain: in these cases, having many annotators will not help to reach a decision and it can only be concluded that it is difficult to classify this part of a text. Different levels of uncertainty could then be introduced, where the absence of a clear majority could be an indication of weak certainty or uncertainty, and a very weak majority could result in an undefined classification.

However, even though different levels of certainty or uncertainty are interesting when studying how uncertainties are expressed and perceived, they would complicate the process of information extraction. Thus, if the final aim of the annotation is to create a system that automatically detects what is certain or uncertain, it would of course be more desirable to have an annotation with a higher inter-annotator agreement. One way of achieving a this would be to provide more detailed annotation guidelines for what to define as certainty and uncertainty. However, when it comes to such a vague concept as uncertainty, there is always a thin line between having guidelines capturing the general perception of uncertainty in the language and capturing a definition of uncertainty that is specific to the writers of the guidelines. Also, there might perhaps be a risk that the complex concept of certainty and uncertainty becomes overly simplified when it has to be formulated as a limited set of guidelines. Therefore, a more feasible method of achieving higher agreement is probably to instead

Class Neg-Spec	Relevant	Retrieved	Corpus	Precision	Recall	F-score
Negation	782	890	853	0.879	0.917	0.897
Speculative words	376	558	1061	0.674	0.354	0.464
Total	1 158	1 448	1 914	0.800	0.605	0.687

Table 3: The results for *negation* and *speculation* on consensus when executing Stanford NER CRF using ten-fold cross validation.

Class Cert-Uncertain	Relevant	Retrieved	Corpus	Precision	Recall	F-score
Certain expression	4 022	4 903	4 745	0.820	0.848	0.835
Uncertain expression	214	433	577	0.494	0.371	0.424
Undefined expression	2	5	144	0.400	0.014	0.027
Total	4 238	5 341	5 466	0.793	0.775	0.784

Table 4: The results for *certain* and *uncertain* on consensus when executing Stanford NER CRF using ten-fold cross validation.

simplify what is being annotated, and not annotate for such a broad concept as uncertainty in general.

Among other suggestions for improving the annotation guidelines for the corpus that the consensus is based on, Velupillai (2010) suggests that the guidelines should also include instructions on the focus of the uncertainties, that is, what concepts are to be annotated for uncertainty.

The task could thus, for example, be tailored towards the information that is to be extracted, and thereby be simplified by only annotating for uncertainty relating to a specific concept. If diseases or symptoms that are present in a patient are to be extracted, the most relevant concept to annotate is whether a finding is present or not present in the patient, or whether it is uncertain if it is present or not. This approach has, for example, achieved a very high inter-annotator agreement in the annotation of the evaluation data used by Chapman et al. (2001). Even though this approach is perhaps linguistically less interesting, not giving any information on uncertainties in general, if the aim is to search for diseases and symptoms in patients, it should be sufficient.

In light of the discussion above, the question to what extent the annotations in the constructed consensus capture a general perception of certainty or uncertainty must be posed. Since it is constructed using a majority method with three annotators, who had a relatively low pairwise agreement, the corpus could probably not be said to be a precise capture of what is a certainty or uncertainty. However, as Artstein and Poesio (2008) point out, it cannot be said that there is a fixed level of agreement that is valid for all purposes of a corpus, but the agreement must be high enough for a certain purpose. Therefore, if the information on whether there was a unanimous annotation of a sentence or not is retained, serving as an indicator of how typical an expression of certainty or uncertainty is, the constructed corpus can be a useful resource. Both for studying how uncertainty in clinical text is constructed and perceived, and as one of the resources that is used for learning to automatically detect certainty and uncertainty in clinical text.

4 Results of training with Stanford NER CRF

As a first indication of whether it is possible to use the annotated consensus corpus for finding negation and speculation in clinical text, we trained the Stanford NER CRF, (Finkel et al., 2005) on the annotated data. Artstein and Poesio (2008) write that the fact that annotated data can be generalized and learnt by a machine learning system is not an indication that the annotations capture some kind of reality. If it would be shown that the constructed consensus is easily generalizable, this can thus not be used as an evidence of its quality. However, if it would be shown that the data obtained by the annotations cannot be learnt by a machine learning system, this can be used as an indication that the data is not easily generalizable and that the task to learn perhaps should, if possible, be simplified. Of course, it could also be an indication that another learning algorithm should be used or other features selected.

We created two training sets of annotated consensus material.

The first training set contained annotations on the second level, thus annotations that contained the classes *Speculative words* and *Negation*. In 76 cases, the tag for *Negation* was inside an annotation for *Speculative words*, and these occurrences

Class Neg-Spec Bio	Relevant	Retrieved	Corpus	Precision	Recall	F-score
Negation	843	864	872	0.976	0.967	0.971
Speculative words	1 021	1 079	1 124	0.946	0.908	0.927
Scope[1]	1 295	1 546	1 595[2]	0.838	0.812	0.825

Table 5: The results for *negations, speculation cues* and *scopes* on the BioScope Corpus when executing Stanford NER CRF using ten-fold cross validation.

Class Neg-Spec	Relevant	Retrieved	Corpus	Precision	Recall	F-score
Negation A	791	1 005	896	0.787	0.883	0.832
Speculative words	684	953	1 699	0.718	0.403	0.516
Negation F	938	1097	1023	0.855	0.916	0.884
Speculative words	464	782	1 496	0.593	0.310	0.407
Negation H	722	955	856	0.756	0.843	0.797
Speculative words	552	853	1 639	0.647	0.336	0.443

Table 6: The results for *negations* and *speculation cues* and *scopes* for annotator A, F and H respectively when executing Stanford NER CRF using ten-fold cross validation.

of the tag *Negation* were removed. It is detecting this difference between a real negation cue and a negation word inside a cue for speculation that is one of the difficulties that distinguishes the learning task from a simple string matching.

The second training set only contained the consensus annotations on the first level, thus the annotation classes *Certain*, *Uncertain* and *Undefined*.

We used the default settings on Stanford NER CRF. The results of the evaluation using ten-fold cross validation (Kohavi, 1995) are shown in Table 3 and Table 4.

As a comparison, and to verify the suitability of the chosen machine learning method, we also trained and evaluated the BioScope Corpus using Stanford NER CRF for negation, speculation and scope. The results can be seen in Table 5. When training the detection of scope, only BioScope sentences that contained an annotation for negation and speculation were selected for the training and evaluation material for the Stanford NER CRF. This division into two training sets follows the method used by Morante and Daelemans (2009), where sentences containing a cue are first detected, and then, among these sentences, the scope of the cue is determined.

We also trained and evaluated the annotations that were carried out by each annotator A, F and H separately, i.e. the source of consensus. The results can be seen in Table 6.

We also compared the distribution of *Negation* and *Speculative words* in the consensus versus the BioScope Corpus and we found that the consensus, in Swedish, used about the same number of (types) for negation as the BioScope Corpus in English (see Table 7), but for *speculative words* the consensus contained many more types than the BioScope Corpus. In the constructed consensus, 72 percent of the *Speculative words* occurred only once, whereas in the BioScope Corpus this was the case for only 24 percent of the *Speculative words*.

Type of word	Cons.	Bio
Unique words (Types) annotated as *Negation*	13	19
Negations that occurred only once	5	10
Unique words (Types) annotated as *Speculative*	408	79
Speculative words that occurred only once	294	19

Table 7: Number of unique words both in the Consensus and in the BioScope Corpus that were annotated as *Negation* and as *Speculative words*, and how many of these that occurred only once.

5 Discussion

The training results using our clinical consensus corpus in Swedish gave a precision of 87.9 percent and a recall of 91.7 percent for negation cues and a precision of 67.4 percent and a recall of 35.4 percent for speculation cues. The results for detecting negation cues are thus much higher than for detecting cues for speculation using Stanford NER CRF. This difference is not very surprising, given

[1]The scopes were trained and evaluated separetely from the negations and speculations.

[2]The original number of annotated scopes in the BioScope Corpus is 1 981. Of these, 386 annotations for nested scopes were removed.

the data in Table 7, which shows that there are only a very limited number of negation cues, whereas there exist over 400 different cue words for speculation. One reason why the F-score for negation cues is not even higher, despite the fact that the number of cues for negations is very limited, could be that a negation word inside a tag for *speculative words* is not counted as a negation cue. Therefore, the word *not* in, for example, *not really* could have been classified as a negation cue by Stanford NER CRF, even though it is a cue for speculation and not for negation. Another reason could be that the word meaning *without* in Swedish (*utan*) also means *but*, which only sometimes makes it a negation cue.

We can also observe in Table 4, that the results for detection of uncertain expressions are very low (F-score 42 percent). For undefined expressions, due to scarce training material, it is not possible to interpret the results. For certain expressions the results are acceptable, but since the instances are in majority, the results are not very useful.

Regarding the BioScope Corpus we can observe (see Table 5) that the training results both for detecting cues for negation and for speculations are very high, with an F-score of 97 and 93 percent, respectively. For scope detection, the result is lower but acceptable, with an F-score of 83 percent. These results indicate that the chosen method is suitable for the learning task.

The main reason for the differences in F-score between the Swedish consensus corpus and the BioScope Corpus, when it comes to the detection of speculation cues, is probably that the variation of words that were annotated as *Speculative word* is much larger in the constructed consensus than in the BioScope Corpus.

As can be seen in Table 7, there are many more types of speculative words in the Swedish consensus than in the BioScope Corpus. We believe that one reason for this difference is that the sentences in the constructed consensus are extracted from a very large number of clinics (several hundred), whereas the BioScope Corpus comes from one radiology clinic. This is supported by the findings of Szarvas (2008), who writes that cues for speculation are domain-specific. In this case, however, the texts are still within the domain of clinical texts.

Another reason for the larger variety of cues for speculation in the Swedish corpus could be that the guidelines for annotating the BioScope Corpus and the method for creating a consensus were different.

When comparing the results for the individual annotators with the constructed consensus, the figures in Tables 3 and 6 indicate that there are no big differences in generalizability. When detecting cues for negation, the precision for the consensus is better than the precision for the individual annotations. However, the results for the recall are only slightly better or equivalent for the consensus than for the individual annotations. If we analyse the speculative cues we can observe that the consensus and the individual annotations have similar results.

The low results for learning to detect cues for speculation also serve as an indicator that the task should be simplified to be more easily generalizable. For example, as previously suggested for increasing the inter-annotator agreement, the task could be tailored towards the specific information that is to be extracted, such as the presence of a disease in a patient.

6 Future work

To further investigate if a machine learning algorithm such as Conditional Random Fields can be used for detecting speculative words, more information needs to be provided for the Conditional Random Fields, such as part of speech or if any of the words in the sentence can be classified as a symptom or a disease. One Conditional Random Fields system that can treat nested annotations is CRF++ (CRF++, 2010). CRF++ is used by several research groups and we are interested in trying it out for the negation and speculation detection as well as scope detection.

7 Conclusion

A consensus clinical corpus was constructed by applying a few basic rules for combining three individual annotations into one. Compared to the individual annotations, the consensus contained fewer annotations of uncertainties and fewer annotations that divided the sentences into clauses. It also contained fewer annotations for speculative words, and more annotations for negations. Of the sentences in the constructed corpus, 92 percent were identically annotated by at least two persons.

In comparison with the BioScope Corpus, the constructed consensus contained both a larger number and a larger variety of speculative cues.

This might be one of the reasons why the results for detecting cues for speculative words using the Stanford NER CRF are much better for the BioScope Corpus than for the constructed consensus corpus; the F-scores are 93 percent versus 46 percent.

Both the BioScope Corpus and the constructed consensus corpus had high values for detection of negation cues, F-scores 97 and 90 percent, respectively.

As is suggested by Velupillai (2010), the guidelines for annotation should include instructions on the focus of the uncertainties. To focus the decision of uncertainty on, for instance, the disease of a patient, might improve both the inter-annotator agreement and the possibility of automatically learning to detect the concept of uncertainty.

Acknowledgments

We are very grateful for the valuable comments by the three anonymous reviewers.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

CRF++. 2010. CRF++: Yet another CRF toolkit, May 8. http://crfpp.sourceforge.net//.

Hercules Dalianis and Sumithra Velupillai. 2010. How certain are clinical assessments? Annotating Swedish clinical text for (un)certainties, speculations and negations. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May.

Peter L. Elkin, Steven H. Brown, Brent A. Bauer, Casey S. Husser, William Carruth, Larry R. Bergstrom, and Dietlind L. Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making*, 5(1):13.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370.

Yang Huang and Henry J. Lowe. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association*, 14(3):304.

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 180–183. Association for Computational Linguistics.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145.

Roser Morante and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 21–29. Association for Computational Linguistics.

Lior Rokach, Roni Romano, and Oded Maimo. 2008. Negation recognition in medical narrative reports. *Information Retrieval*, 11(6):499–538.

Maria Skeppstedt. 2010. Negation detection in Swedish clinical text. In *Louhi'10 - Second Louhi Workshop on Text and Data Mining of Health Documents, held in conjunction with NAACL HLT 2010*, Los Angeles, June.

SNOMED-CT. 2010. Systematized nomenclature of medicine-clinical terms, May 8. http://www.ihtsdo.org/snomed-ct/.

Stefan Steidl, Michael Levit, Anton Batliner, Elmar Nöth, and Heinrich Niemann. 2005. "Off all things the measure is man" Automatic classification of emotions and inter-labeler consistency. In *Proceeding of the IEEE ICASSP,2005*, pages 317–320.

György Szarvas. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08: HLT*, pages 281–289, Columbus, Ohio, June. Association for Computational Linguistics.

Sumithra Velupillai. 2010. Towards a better understanding of uncertainties and speculations in swedish clinical text – analysis of an initial annotation trial. To be published in the proceedings of the Negation and Speculation in Natural Language Processing Workshop, July 10, 2010, Uppsala, Sweden.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S-11).