Speculation and negation annotation in natural language texts: what the case of BioScope might (not) reveal

Veronika Vincze
University of Szeged
Szeged, Hungary
vinczev@inf.u-szeged.hu

1 Introduction

In information extraction, it is of key importance to distinguish between facts and uncertain or negated information. In other words, IE applications have to treat sentences / clauses containing uncertain or negated information differently from factual information that is why the development of hedge and negation detection systems has received much interest – e.g. the objective of the CoNLL-2010 Shared Task was also to develop hedge detection systems (Farkas et al., 2010). For the training and evaluation of such systems, corpora annotated for negation and speculation are necessary.

There are several linguistic phenomena that can be grouped under the term uncertainty. Besides hedge and speculation, doubtful events are also considered as a subtype of uncertainty (Kim et al., 2008) and Ganter and Strube (2009) argue that the notion of *weasel words* are similar to hedges. A word is considered to be a weasel word if it creates an impression that something important has been said, but what is really communicated is vague, misleading, evasive or ambiguous, thus, it is also related to uncertainty. All these phenomena might be of interest for IE applications, which yields that the creation of corpora with uncertainty annotation is indispensable.

2 Related work

There exist some corpora that contain annotation for speculation and/or negation. The GENIA Event corpus (Kim et al., 2008) annotates biological events with negation and two types of uncertainty. In the BioInfer corpus (Pyysalo et al., 2007) biological relations are annotated for negation. The system developed by Medlock and Briscoe (2007) made use of a corpus consisting of six papers from genomics literature in which sentences were annotated for speculation. Settles et al. (2008) constructed a corpus where sentences are classified as either speculative or definite, however, no keywords are marked in the corpus and Shatkay et al. (2008) describe a database where sentences are annotated for certainty among other features. As a corpus specifically annotated for weasel words, WikiWeasel should be mentioned, which was constructed for the CoNLL-2010 Shared Task (Farkas et al., 2010) and contains Wikipedia paragraphs annotated for weasel words.

3 The BioScope corpus

The BioScope corpus (Vincze et al., 2008) is – to our best knowledge – the largest corpus available that is annotated for both negation and hedge keywords and the only one that contains annotation for linguistic scopes. It includes three types of texts from the biomedical domain – namely, radiological reports, biological full papers and abstracts from the GENIA corpus. (15 new full biomedical papers were annotated for hedge cues and their scopes, which served as the evaluation database of the CoNLL-2010 Shared Task (Farkas et al., 2010), and this dataset will be added to BioScope in the near future.) The annotation was carried out by two students of linguistics supervised by a linguist. Problematic cases were continuously discussed among the annotators and dissimilar annotations were later resolved by the linguist.

3.1 Annotation principles

In BioScope, speculation is understood as the possible existence of a thing is claimed – neither its existence nor its non-existence is known for sure. Only one level of uncertainty is marked (as opposed to the GENIA corpus (Kim et al., 2008) or Shatkay et al. (2008)) and no weasels are annotated. Negation is seen as the implication of non-existence of something.

The annotation was based on four basic principles:

Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 28–31,
Uppsala, July 2010.

- Each keyword has a scope.
- The scope must include its keyword.
- Min-max strategy:
  - The minimal unit expressing hedge/negation is marked as keyword.
  - The scope is extended to the maximal syntactic unit.
- No intersecting scopes are allowed.

These principles were determined at the very beginning of the annotation process and they were strictly followed throughout the corpus building.

3.2 Problematic cases

However, in some cases, some language phenomena seemed to contradict the above principles. These issues required a thorough consideration of the possible solutions in accordance with the basic principles in order to keep the annotation of the corpus as consistent as possible. The most notable examples include the following:

- Negative keywords without scope:

  > [Negative] chest radiograph.

  In this case, the scope contains only the keyword.

- Elliptic sentences

  > Moreover, ANG II stimulated NF-kappaB activation in human monocytes, but [not] in lymphocytes from the same preparation.

  With the present encoding scheme of scopes, there is no way to signal that the negation should be extended to the verb and the object as well.

- Nested scopes

  One scope includes another one:

  > These observations (suggest that TNF and PMA do (not lead to NF-kappa B activation through induction of changes in the cell redox status)).

  The semantic interpretation of such nested scopes should be understood as ”it is possible that there is no such an event that. . . ”.

- Elements in between keyword and target word

  Although *however* is not affected by the hedge cue in the following example, it is included in the scope since consecutive text spans are annotated as scopes:

  > (Atelectasis in the right mid zone is, however, <possible>).

- Complex keywords

  Sometimes a hedge / negation is expressed via a phrase rather than a single word: these are marked as complex keywords.

- Inclusion of modifiers and adjuncts

  It is often hard to decide whether a modifier or adjunct belongs to the scope or not. In order not to lose potentially important information, the widest scope possible is marked in each case.

- Intersecting scopes

  When two keywords occur within one sentence, their scopes might intersect, yielding one apparently empty scope (i.e. scope without keyword) and a scope with two keywords:

  > (Repression did ([not] <seem> to involve another factor whose activity is affected by the NSAIDs)).

  In such cases, one of the scopes (usually the negative one) was extended:

  > ((Repression did [not] <seem> to involve another factor whose activity is affected by the NSAIDs)).

On the other hand, there were some cases where the difficulty of annotation could be traced back to lexical issues. Some of the keyword candidates have several senses (e.g. *if*) or can be used in different grammatical structures (e.g. *indicate* vs. *indicate that*) and not all of them are to be marked as a keyword in the corpus. Thus, senses / usages to be annotated and those not to be annotated had to be determined precisely.

Finally, sometimes an apparently negative keyword formed part of a complex hedge keyword (e.g. *cannot be excluded*), which refers to the fact that speculation can be expressed also by a negated word, thus, the presence of a negative word does not automatically entail that the sentence is negated.

4 Outlook: Comparison with other corpora

Besides BioScope, the GENIA Event corpus (Kim et al., 2008) also contains annotation for negation and speculation. In order to see what the main differences are between the corpora, the annotation principles were contrasted:

- in GENIA Event, no modifier keywords are marked, however, in BioScope, they are;
- the scope of speculation and negation is explicitly marked in BioScope and it can be extended to various constituents within the clause / sentence though in GENIA Event, it is the event itself that is within the scope;
- two subtypes of uncertainty are distinguished in GENIA Event: *doubtful* and *probable*, however, in BioScope there is one umbrella term for them (*speculation*).

An essential difference in annotation principles between the two corpora is that GENIA Event follows the principles of event-centered annotation while BioScope annotation does not put special emphasis on events. Event-centered annotation means that annotators are required to identify as many events as possible within the sentence then label each separately for negation / speculation.

The multiplicity of events in GENIA and the maximum scope principle exploited in BioScope (see 3.1) taken together often yields that a GENIA event falls within the scope of a BioScope keyword, however, it should not be seen as a speculated or negated event on its own. Here we provide an illustrative example:

> In summary, our data suggest that changes in the composition of transcription factor AP-1 is a key molecular mechanism for increasing IL-2 transcription and may underlie the phenomenon of costimulation by EC.

According to the BioScope analysis of the sentence, the scope of *suggest* extends to the end of the sentence. It entails that in GENIA it is only the events *is a key molecular mechanism* and *underlie the phenomenon* that are marked as probable, nevertheless, the events *changes*, *increasing*, *transcription* and *costimulation* are also included in the BioScope speculative scope. Thus, within this sentence, there are six GENIA events out of which two are labeled as probable, however, in BioScope, all six are within a speculative scope.

In some cases, there is a difference in between what is seen as speculative / negated in the corpora. For instance, negated ”investigation” verbs in Present Perfect are seen as doubtful events in GENIA and as negative events in BioScope:

> However, a role for NF-kappaB in human CD34(+) bone marrow cells has not been described.

According to GENIA annotation principles, the role has not been described, therefore it is doubtful what the role exactly is. However, in BioScope, the interpretation of the sentence is that there has not been such an event that the role for NF-kappaB in human CD34(+) bone marrow cells has been described. Thus, it is marked as negative.

Another difference between the annotation schemes of BioScope and GENIA is that instances of weaseling are annotated as probable events in GENIA, however, in BioScope they are not. An example for a weasel sentence is shown below:

> Receptors for leukocyte chemoattractants, including chemokines, are traditionally considered to be responsible for the activation of special leukocyte functions such as chemotaxis, degranulation, and the release of superoxide anions.

5 Conclusions

Some interesting conclusions can be drawn from the difficulties encountered during annotation process of the BioScope corpus. As for methodology, it is unquestionable that precisely defined rules (on scope marking, keyword marking and on the interpretation of speculation / negation) are essential for consistent annotation, thus, pre-defined guidelines can help annotation work a lot. However, difficulties or ambiguities not seen previously may emerge (and they really do) only during the process of annotation. In this way, a continuous reformulation and extension of annotation rules is required based on the corpus data. On the other hand, problematic issues sometimes might be solved in several different ways. When deciding on their final treatment, an ideal balance between gain and loss should be reached, in other words, the min-max strategy as a basic annotation

principle can also be applied here (minimize the loss and maximize the gain that the solution can provide).

Acknowledgments

This work was supported in part by the National Office for Research and Technology (NKTH, `http://www.nkth.gov.hu/`) of the Hungarian government within the framework of the project MASZEKER.

References

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

Viola Ganter and Michael Strube. 2009. Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 173–176, Suntec, Singapore, August. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(Suppl 10).

Ben Medlock and Ted Briscoe. 2007. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the ACL*, pages 992–999, Prague, Czech Republic, June.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*, pages 1–10.

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

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(Suppl 11):S9.