Annotating Negation and Speculation: Annotation Guidelines

Natalia Konstantinova
Sheila C. M. de Sousa
Research Group in Computational Linguistics
University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB, UK
{n.konstantinova,sheila.castilhomonteirodesousa}@wlv.ac.uk

Contents

This document presents the guidelines that can be used for the annotation of speculation and negation in the review domain. The guidelines are split into two parts: speculation and negation. Each part provides information about the marking schemes, the keywords used and the scopes to be annotated.

1 Terminology

Speculation and negation are important aspects of language. **Speculation** is related to the broader concept of "modality" which has been extensively studied both in linguistics and philosophy (Saurí, 2008). Various classifications of modality can be found in literature (Morante and Daelemans, 2009). Related terms like "hedging", "evidentiality", "uncertainty", and "factuality" are also used when talking about different aspects of modality. Saurí et al. (2006) state that modality "expresses the speaker's degree of commitment to the events being referred to in a text".

Negation is part of the broader concept of "polarity", which indicates whether a statement is presented as positive or negative (Saurí, 2008). In simple propositional logic, negation is an operator that reverses the truth value of a proposition (Miestamo, 2007).

In defining speculation and negation we follow the definitions introduced by Vincze (2010):

- "speculation is understood as the possible existence of a thing is claimed – neither its existence nor its non-existence is known for sure", so there is not enough evidence in the text to say whether information is true or not.
- "negation is seen as the implication of non-existence of something".

These two phenomena are interrelated (de Haan, 1997) and have similar characteristics in the text: they both have **scope**, so affect the part of the text which is denoted by the presence of negation or speculation cue words.

2 General remarks

There are several general principles to be followed when annotating negation and speculation:

- Only sentences with some instance of speculative language or negation should be considered.

- Questions should not be annotated for negation, nor for speculation.
- **Min-max strategy** should be followed during the annotation:
  - When annotating keywords, try to choose **the minimal unit** which expresses negation or speculation (special attention should be paid to distinguishing **complex cues** and sequences of several keywords) (these will be discussed in more detail later)
  - When annotating scope, try to annotate **the maximum words** affected by the phenomenon:

    [1] *They ended up hitting me in the nuts, which, to say the least, was* ***probably***$_{spec}$ *(better than what the director of this film did to the memory of Dr.Seuss).*

    In this example the scope of the cue word *probably* includes all the words after it.

    [2] *I* ***'d***$_{spec}$ *(rather buy a set at target and replace it every year for the convenience of being able to put them in the dishwasher and their lighter construction).*

    Even though the scope of *'d* can seem long, it includes exactly the part of the sentence effected by the speculation phenomenon. Therefore the scope includes not only verbal phrases, but also adverbial modifiers.
- The cue words are not included in the scope:

  [3] ***Maybe***$_{spec}$ *(the patients that say they are hearing voices really are hearing voices).*
- Transitional words (e.g *in addition*, *not to mention* etc.) should not be included in the scope:

  [4] *I* ***think***$_{spec}$*, however, (it was his mistake to go there alone).*
- When unsure of the scope, annotate only a keyword.
- When unsure what type the keyword should be assigned to (whether it expresses negation or speculation), use **'undecided'** category.

3 Speculation

Speculative elements are marked with the tag 'CUE' which is described by 'ID' and 'TYPE' (type equals to 'speculation' in this case). Their scope is marked with the tag 'XCOPE'. The link between the scope and the ID word is marked with the tag 'REF', where 'SRC' is the ID of the cue it belongs to.

To illustrate examples of the annotation process we use the keywords in bold and their types in subscript; we use () to indicate the boundaries of the scope of speculative keywords:

[5] *(The Casino)* ***seemed***$_{spec}$ *(to be like any other Casino in Vegas).*

3.1 Keywords

The most typical instances of speculative language, that is, keywords, can be grouped as follows:

1. **auxiliaries:** may, might, can, would, should, could.
2. **verbs with speculative content:** think, suggest, question, presume, suspect, indicate, suppose, seem, appear, expect, etc.
3. **adjectives or adverbs:** probable, likely, possible, apparently, unsure, etc.
4. **conjunctions:** if, or, and/or, either ... or, versus, vs., etc.

Given that we are dealing with the review domain, cases of **ungrammatical use** of the language are quite common, so ungrammatical variants of the keywords should be considered as well.

Complex keywords: When speculation or negation is expressed through a phrase rather than a single word and these words cannot express speculation separately, they are annotated as complex keywords:

[6] *I* ***have a feeling***$_{spec}$ *(that many readers would have given up before the end due to boredom, frustration or the maddening feeling of ' What the hell is Patterson thinking when he wrote this?').*

In this case, *have a feeling* could be substituted by (I) *think* which clearly expresses uncertainty. However the words *have*, *a*, *feeling*, *that* cannot express uncertainty on their own.

Complex keywords should be carefully distinguished from **the sequence** of keywords. Consider an example below:

[7] *I* ***didn't***$_{neg}$ *[****think***$_{spec}$ *(it* ***would***$_{spec}$ *(be* ***possible***$_{spec}$ *(for anyone to rip the heart out of a Dr. Seuss book)))].*

In this example we have four separate cue words and they do not combine to become complex cue words. All of them can express negation or speculation on their own, so they are treated as a sequence of keywords.

[8] *The Altima* ***would***$_{spec}$ *(****probably***$_{spec}$ *(be a very acceptable package priced at say 19-22k max.))*

The keywords *would* and *probably* also do not form a complex keyword and should be annotated as separate items.

Minimal units expressing speculation should be annotated as a keyword, therefore prepositions, determiners, adverbs, etc. should not be annotated as a part of a keyword:

[9] *I* ***didn't***$_{neg}$ *[****believe***$_{spec}$ *(in this strategy)].*

However, if **the sequence of words** expresses speculation only together, these words should be annotated as complex keyword:

[10] *(It)* ***feels like***$_{spec}$ *(he forgot his words).*

Be careful with **different meanings** of the words - some meanings may not express speculation.

3.2 Scope

As mentioned above, annotate the maximal part of the text affected by the phenomena.

The scope of **verbs, auxiliaries, adjectives and adverbs** usually starts right after the keyword. In the case of verbal elements, i.e. **verbs and auxiliaries**, it ends at the end of the clause (if the verbal element is within a relative clause or a coordinated clause) or sentence, thus, all complements and adjuncts are included.

[11] *You* ***can***$_{spec}$ *(go to the observation deck for free* ***if***$_{spec}$ *(you eat there)).*

[12] *So I* ***suppose***$_{spec}$ *(I just took a lateral step and went for the Stainless Set).*

Attributive and predicative adjectives: The scope of **attributive adjectives** generally extends to the following noun phrase, whereas the scope of **predicative adjectives** includes the whole sentence.

[**13**] *This movie had two* ***possible***$_{spec}$ *(directors) and I'm sure they did the right choice.*

[**14**] *(The end like this is)* ***possible***$_{spec}$*.*

Adverbs: In the cases of both sentential and other adverbs the scope usually starts right after the keyword.

[**15**] *I am* ***probably***$_{spec}$ *(skipping this part of the book), as there is nothing interesting there.*

[**16**] *The same type of person who* ***probably***$_{spec}$ *(pays someone else to wash the dishes).*

Conjunctions: In the case of the review domain, as the keywords were not included in the scope, the scopes were annotated separately and then linked to the keywords:

[**17**] *As far as I remember , vacation with accommodation in (Rio), (Golden Nugget),(Excalibur)* ***or***$_{spec}$ *(Las Vegas Hilton) were available for cheaper rates than what I paid for Riviera.*

[**18**] *The book is full of (inconsistencies)* ***or***$_{spec}$ *(muddy areas).*

[**19**] *Either ((that))* ***or***$_{spec}$ *((he was specifically looking for a movie contract for this story)).*

The scope of the **'either...or'** is linked to both keywords.

Passive voice:

It should be noted that unlike active voice sentences passive ones include subject into the scope of the keyword:

[**20**] *(He)****was presumed***$_{spec}$ *(to be a good choice at the beginning).*

[**21**] *(This song)* ***is considered***$_{spec}$ *(a classic by all) .*

In order to check what parts of the passive sentence should be included into the scope, paraphrase the sentence into a non-passive sentence and see what will be the scope. Then annotate in a similar way in passive voice.

So the previous examples can be rephrased in the following way:

[**22**] *It was* ***presumed***$_{spec}$ *(that he was a good choice at the beginning).*

[23] *It* ***is considered***$_{spec}$ *(that this song is a classic by all) .*

Special attention should be paid to **'seem'** (same with **'appear'**) and two possible constructions it can participate in should be distinguished:

[24] *It* ***seems***$_{spec}$ *(this movie was a great success).*

[25] *(He)* ***seems***$_{spec}$ *(to know his job really well).*

[26] *(The motherboard)* ***appears***$_{spec}$ *(to be an Intel Motherboard, consistent with most Dell configurations).*

Embedded scopes: Although keywords are not included in their own scope, a keyword can be included in the scope of other keywords and situations of embedded scopes are possible:

[27] *I'm* ***not sure***$_{spec}$ *(****if***$_{spec}$ *(he* ***should***$_{spec}$ *((be angrier at his widow for giving studios the rights to his stories),* ***or***$_{spec}$ *(to the studios for stabbing his widow in the back when she trusted them))))).*

[28] *I* ***think***$_{spec}$ *(it* ***might***$_{spec}$ *(be nice to have a few pieces for certain dishes)).*

No scope: Unlike the BioScope guidelines which mention only the cases of negation keywords without scope, situations where speculation keywords had no scope were encountered as well in the review domain:

[29] *This movie didn't have anything to do with a children's movie as it* ***should***$_{spec}$*.*

Here we can see the case of ellipsis which results in the keyword having no scope.

Not sure: Also it was noted that the case of the keyword *not sure* can be difficult for annotation as its scope should include all the elements it modifies, for instance, it should include all the elements on the right in the following example:

[30] ***not sure***$_{spec}$ *(if he should be angrier at his widow for giving studios the rights to his stories, or to the studios for stabbing his widow in the back when she trusted them).*

4 Negation

Negative elements are marked with the tag 'CUE' which is described by 'ID' and 'TYPE' (type equals to 'negation' in this case). Their scope is marked with the tag 'XCOPE'. The link between the scope and the ID word is marked with the tag 'REF', where 'SRC' is the ID of the cue it belongs to.

To illustrate examples of the annotation process we use the keywords in bold and their types in subscript; and [] to indicate the scope of negative keywords:

[31] *This is the writing of a complete amateur,* ***not***$_{neg}$ *[someone who is at the top of his game].*

4.1 Keywords

The most typical instances of negative keywords are listed here:

1. auxiliary: cannot
2. adjectives or adverbs: impossible, impossibly etc.
3. conjunctions: neither ... nor etc.
4. negation words: no, not
5. preposition: without

As it was mentioned before, sentences including a negative keyword are not necessarily to be annotated for negation. They can, however, have **speculative content** as well:

[32] ***not sure***$_{spec}$ *(it was the best idea to stay till the end).*

Given that we are dealing with the review domain, cases of **ungrammatical use** of the language are quite common, so ungrammatical variants of the keywords should be considered as well.

[33] *A*nd she **ain't**$_{neg}$ [**no**$_{neg}$ [Rossellini]].

Unlike BioScope corpus we annotate cases similar to 'almost no' and 'not always' (which are ignored there), because we think that for further processing it is important to capture these cases as well:

[34] *The film had almost* ***no***$_{neg}$ *[relation to the book].*

[35] *That's right,* ***no***$_{neg}$ *[more burnt popcorn] and almost* ***no***$_{neg}$ *[more unpopped kernels].*

[36] *Bigger amount of money invested in the development is* ***not***$_{neg}$ *[always beneficial for the final product].*

4.2 Scope

Negation scope: Similar to the BioScope guidelines for the negation scope, only the words that are modified by the negation cue are included in the scope:

[37] *It* ***isn't***$_{neg}$ *[scary], but it is enthralling.*

The scope of **negative auxiliaries, adjectives and adverbs** usually starts right with the keyword and ends at the end of the phrase, clause or sentence:

[38] *It is* ***impossible***$_{neg}$ *[to see the difference between the two models].*

When the subject of a passive or active sentence contains the negative determiner **no**, its scope extends to the entire sentence:

[39] ***No***$_{neg}$ *[phone was damaged during our tests].*

Negative conjunctions generally have scope over the syntactic unit whose members it coordinates.

Complex keywords have got one scope:

[40] *The Taurus handles moderately well for a mid-sized family sedan; cornering is responsive with* ***neither***$_{neg}$ *[crispness]* ***nor***$_{neg}$ *[poor control].*

The scope of the **'neither...nor'** is linked to both keywords.

Passive voice:
Paraphrasing the sentence in active voice helps to identify the correct scope of the negation in passive voice:

[41] *[This book] wasn't [published before the end of 2000].*

can be paraphrase into:

[42] *They* ***didn't***$_{neg}$ *[publish this book before the end of 2000].*

Prepositions have scope over the following (noun) phrase:

[43] *This camera works much better* ***without***$_{spec}$ *[flash].*

Elliptic sentences: For elliptic sentences the keyword is marked and the scope is neglected. When annotating the SFU Review corpus we follow the strategy suggested in the BioScope guidelines:

[**44**] *I later discovered that my 11 year old understood all of them. I wish he* ***hadn't***$_{neg}$.

[**45**] *It has so many features that other laptops do* ***not***$_{neg}$.

There can be also cases when the combination of different types of keywords (ie. negation and speculation ones) results in the **embedded scopes**:

[**46**] It **isn't**$_{neg}$ [(vulgar) **or**$_{spec}$ (sexual)]

5 Difficult cases - more examples

Keyword sequences: The presence of the sequences of the keywords can create additional difficulties for the annotation. The nature of the review domain texts introduces a greater possibility of encountering such cases than e.g. in the biomedical domain. Therefore special care should be taken when distinguishing several keywords that go one after another. Although some examples of two or more keywords in a sequence could be also considered as complex keywords they should be annotated separately if they can express hedge on their own:

[**47**] *I* ***didn't***$_{neg}$ *[****think***$_{spec}$ *(it* ***would***$_{spec}$ *(be* ***possible***$_{spec}$ *(for anyone to rip the heart out of a Dr. Seuss book)))].*

In this example the keywords *didn't* and *think* may seem complex keywords but they should be annotated as separate keywords since *didn't* negates *think* which is the leading cue of the whole idea of speculation.

Great number of keywords: Close attention should be paid to sentences with a great number of keywords, which can lead the annotator to make mistakes. One of these difficult cases is presented below as an illustration:

[**48**] *This creative re-engineering draws (the viewer)*^{1} ***or***$_{1\,spec}$ *(reader)*^{1} *into a parallel universe where age-old lessons* ***can***$_{spec}$ *((be taught)*^{2} ***or***$_{2\,spec}$ *(re-taught)*^{2}*)* ***without***$_{neg}$ *[(the obstructions created in the minds)*^{3,4,5}*,* ***or***$_{3\,spec}$ *(interferences)*^{3,4,5}*,* ***or***$_{4\,spec}$ *(misconceptions)*^{3,4,5} ***if***$_{spec}$ *(you prefer),* ***or***$_{5\,spec}$ *even (pre-concepts)*^{3,4,5}*] that* ***may***$_{spec}$ *(****probably***$_{spec}$ *(lead to misunderstandings)).*

While for the keywords $\mathbf{or}_{1\,spec}$ and $\mathbf{or}_{2\,spec}$ the scopes are easily identified, for the $\mathbf{or}_{3,4,5\,spec}$ the scopes are tricky since they should include all the members modified by the keyword *not* even if these members are syntactically distant from the keywords.

6 Bibliography

References

de Haan, Ferdinand. 1997. *The interaction of modality and negation: a typological study.* Garland Publishing, New York, USA.

Miestamo, Matti. 2007. Negation an overview of typological research. *Language and Linguistics Compass*, 1(5):552–570, September.

Morante, Roser and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saurí, Roser. 2008. *A factuality profiler for eventualities in text.* Ph.D. thesis, Waltham, MA, USA.

Saurí, Roser, Marc Verhagen, and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *In The 19th International FLAIRS Conference, FLAIRS 2006.*

Vincze, Veronika. 2010. Speculation and negation annotation in natural language texts: what the case of BioScope might (not) reveal. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 28–31, Uppsala, Sweden, July. University of Antwerp.