Department of Computer and Systems Sciences

Stockholm University

Annotating named entities in clinical text by combining pre-annotation and active learning

Problems

Potential problems with pre-tagged data:

1. The annotator might be biased to choose the annotation provided by the pre-tagger.
2. If the pre-tagger produces poor pre-taggings on the data given to the annotator, or If there are many possible pre-annotations to choose from, the annotation work is not reduced.

1. Reduce bias

To reduce the bias problem, it is proposed that the two best taggings produced by a pre-tagger are presented to the human annotator, without informing the annotator which of them that the pre-tagger considers most likely.

2. Reduce annotation work

To reduce the annotation work, the instances, for which none of the two presented pre-taggings are considered correct by the annotator must be minimised. To minimise these instances, a version of active learning is proposed, in which text passages are actively selected so that it is likely that one of the two best pre-taggings is correct. A challenge of this approach is to select text passages to present to the annotator that are informative enough to be useful to the learner and for which the pre-tagger is certain enough to produce a maximum of two plausible pre-taggings.

Method

The following proposed method combines pre-tagging with a version of active learning:

1) Train machine learning model with annotated data.

2) Apply machine learning model on un-annotated data. Let the model provide the three most probable pre-taggings for each sentence in the data, together with their level of certainty.

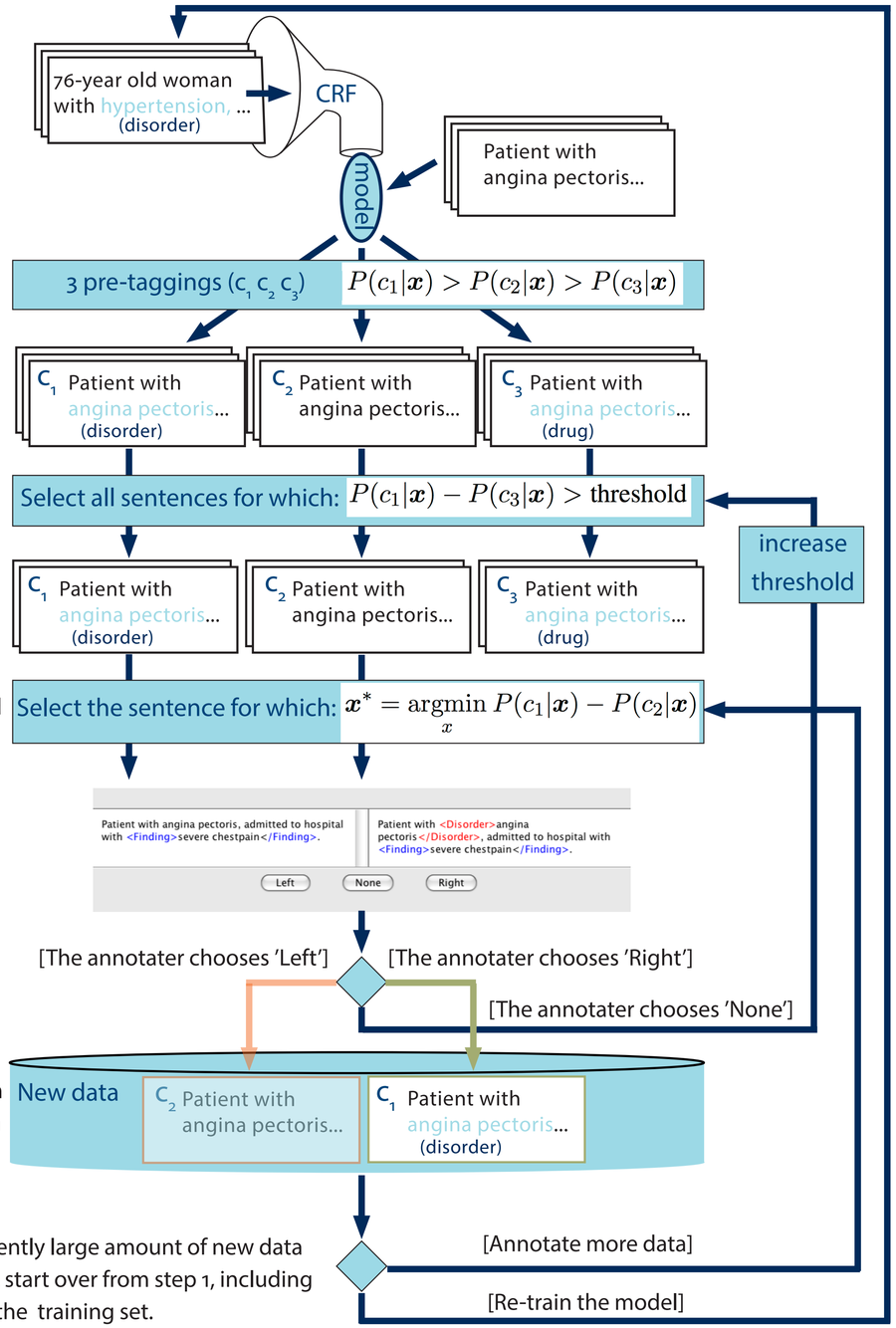
3) Only retain the sentences for which the two best suggestions are much better than the third. Given by a confidence threshold, initially set to zero.

4) Select the sentence with the lowest confidence difference between the best and the second best pre-tagging.

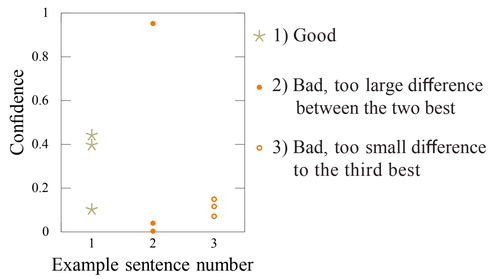
5) Present the two alternative pre-taggings without revealing which is given the highest confidence. Let the annotator choose that the right, the left or none of the pre-taggings is correct.

6) If 'None' is selected, increase the threshold, and start over from step 3. Else, add the selected data to the set of un-annotated data and continue from step 4.

7) When a sufficiently large amount of new data has been added, start over from step 1, including the new data in the training set.

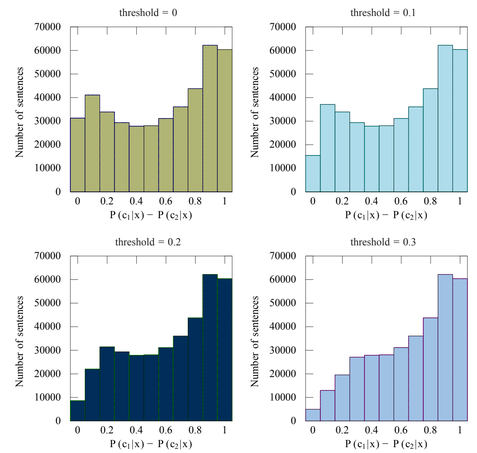


Three example sentences



Experiments

A CRF++ model was trained on clinical text annotated for the four entities: Disorder, Finding, Pharmaceutical drug and Body structure. This model was thereafter applied to texts from other clinical domains. The difference in confidence between the two best pre-taggings was measured for four different thresholds for distance to confidence for the third best pre-tagging.



Maria Skeppstedt

(mariask@dsv.su.se)