A Simple Ensemble Method for Hedge Identification

Ferenc P. Szidarovszky[1], **Illés Solt**[1], **Domonkos Tikk**[1,2]
[1] Budapest University of Technology and Economics, Budapest, Hungary
[2] Humboldt-Universität zu Berlin, Berlin, Germany
ferenc.szidarovszky@hotmail.com,{solt,tikk}@tmit.bme.hu

Abstract

We present in this paper a simple hedge identification method and its application on biomedical text. The problem at hand is a subtask of CoNLL-2010 shared task. Our solution consists of two classifiers, a statistical one and a CRF model, and a simple combination schema that combines their predictions. We report in detail on each component of our system and discuss the results. We also show that a more sophisticated combination schema could improve the F-score significantly.

1 Problem definition

The CoNLL-2010 Shared Task focused on the identification and localization of uncertain information and its scope in text. In the first task, a binary classification of sentences had to be performed, based on whether they are uncertain or not. The second task concentrated on the identification of the source of uncertainty – specifying the keyword/phrase that makes its context uncertain –, and the localization of its scope. The organizers provided training data from two application domains: biomedical texts and Wikipedia articles. For more details see the overview paper by the organizers (Farkas et al., 2010). We focused on task 1 and worked with biomedical texts exclusively.

The biomedical training corpus contains selected abstracts and full text articles from the BioScope corpus (Vincze et al., 2008). The corpus was manually annotated for *hedge cues* on the phrase level. Sentences containing at least one cue are considered as uncertain, while sentences with no cues are considered as factual. Though cue tagging was given in the training data, their marking in the submission was not mandatory.

The evaluation of systems at task 1 was performed on the sentence level with the F-measure of the uncertain class being the official evaluation metric. For evaluation, corpora also from both domains were provided that allowed for in-domain and cross-domain experiments as well. Nevertheless, we restricted the scope of our system to the in-domain biomedical subtask.

2 Background

Automatic information extraction methods may incorrectly extract facts that are mentioned in a negated or speculative context. If aiming at high accuracy, it is therefore crucial to be able to classify assertions to avoid such false positives. The importance of assertion classification has been recently recognized by the text mining community, which yielded several text-mining challenges covering this task. For example, the main task of Obesity Challenge (Uzuner, 2008) was to identify based on a free text medical record whether a patient is known to, speculated to or known not to have a disease; in the BioNLP '09 Shared Task (Kim et al., 2009), mentions of bio-molecular events had to be classified as either positive or negative statements or speculations.

Approaches to tackle assertion classification can be roughly organized into following classes: rule based models (Chapman et al., 2001), statistical models (Szarvas, 2008), machine learning (Medlock and Briscoe, 2007), though most contributions can be seen as a combination of these (Uzuner et al., 2009). Even when classifying sentences, the most common approach is to look for cues below the sentence-level (Özgür and Radev, 2009). The common in these approaches is that they use a text representation richer than bag-of-words, usually tokens from a fixed-width window with additional surface features.

Evaluation of assertion classification is mostly performed at the sentence level, where state-of-the-art systems have been reported to achieve an F-measure of 83–85% for hedge detection in

Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task, pages 144–147,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

biomedical literature (Medlock and Briscoe, 2007; Szarvas, 2008).

3 Methods

Although the problem itself is a binary categorization problem, we approach the problem at the token/phrase level. We search for hedge cues and used the decision model also applied by the annotators of the training corpus: when a sentence contains at least one uncertainty cue then it is uncertain, otherwise factual.

We applied two different models to identify hedge cues:

- a *statistical model* that creates a candidate list of cue words/phrases from the training samples, and cuts off the list based on the precision measured on the trial set;
- a sequence tagger *CRF model*, trained again with hedge cues using various feature sets.

Finally, we combined the outputs of the methods at the sentence level. Here we applied two very simple ways of combination: the aggressive one assigns a sentence to the uncertain class if any of the models finds a cue phrase therein (OR merger), while the conservative only if both models predict the sentence as uncertain (AND merger). We submitted the version which produced better result on the trial set. The overview of our system is depicted on Figure 1.

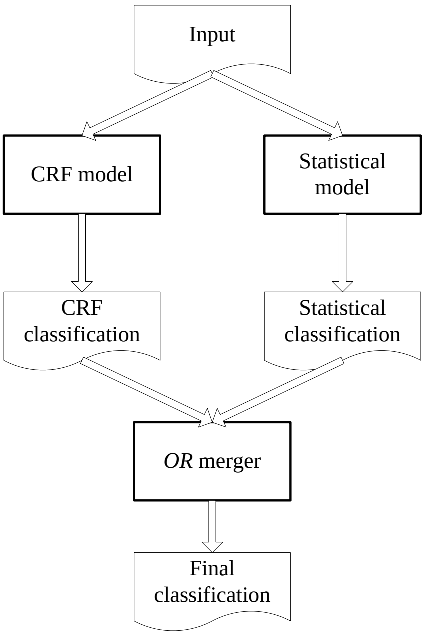

Figure 1: System overview

3.1 Preprocessing

The biomedical corpus was provided in two train/trial pairs (abstracts and full texts), see also Table 1. Because the ratio of uncertain sentences is similar in both train and trial sets, we merged the two train sets and the two trial sets, respectively, to obtain a single train/trial pair. Since the trial set was originally included also in the train set, we removed the elements of the merged trial set from the merged train set. In the following, we refer to them as *train* and *trial* sets. All data (train, trial, evaluation) were given as separate sentences; therefore no sentence segmentation had to be performed.

Merging train and trial sets was also motivated by the sparsity of data and the massively different train/trial ratio observed for the two types of biomedical texts (Table 1). Therefore building separate models for abstracts and full texts may only yield overfitting, particularly because such a distinction is not available for the evaluation set.

3.2 Statistical model

The statistical model considers a sentence uncertain, if it contains at least one cue from a validated set of cue phrases. To determine the set of cue phrases to be used, we first collected all annotated cues from the training data. From this candidate cue set we retained those ones that had a precision over a predefined threshold. To this end we measured on the training set the precision of each cue phrase. We depicted on Figure 2 the precision, recall and F-measure values obtained on the trial set with different cue phrase precision thresholds.

The candidate cue set contains 186 cue phrases, among which 83 has precision 1.0 and 141 has precision greater or equal 0.5. Best cue phrases include words/phrases like *cannot* + *verb phrase, hypothesis, indicate, may, no(t)* + *verb/noun, raise the* + *noun, seem, suggest, whether* etc., while low precision cues are, e.g., *assume, not fully understood, not, or, prediction, likelihood.*

3.3 CRF model

Identifying entities such as speculation cues can be efficiently solved by training conditional random field (CRF) models. As a general sequence tagger, a CRF can be naturally extended to incorporate token features and features of neighboring tokens. The trained CRF model is then applied to unseen

	Train set			Trial set			Evaluation set		
	sentences	uncertain	ratio	sentences	uncertain	ratio	sentences	uncertain	ratio
Abstract	11 832	2 091	17.7 %	39	10	25.6 %	–	–	–
Full text	2 442	468	19.2 %	228	51	22.4 %	–	–	–
Total	14 274	2 559	17.9 %	267	61	22.9 %	5 003	790	15.8 %

Table 1: Basic statistics of the provided train, trial, and evaluation sets

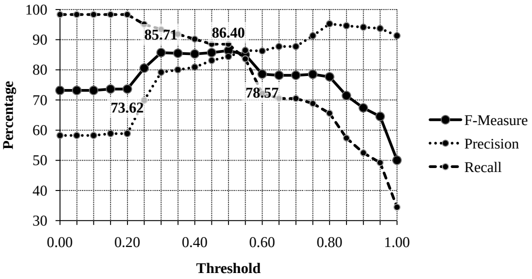

Figure 2: Cue phrase threshold selection

text, whenever a speculation cue is found the containing sentence is annotated as being speculative. In our experiments, we used MALLET (McCallum, 2002) to train CRF models using custom tokenization (Section 3.3.1) and feature sets (Section 3.3.2). We included features of 2–2 neighboring tokens in each direction, not surpassing the sentence limits.

3.3.1 Tokenization

We split text into tokens using punctuation and white-space tokenization, keeping punctuation symbols as separate tokens.

3.3.2 Feature sets

We experimented with the following binary surface features:

1. token text
2. token text in lowercase
3. stem of token in lowercase
4. indicator of the token being all lowercase
5. indicator whether the token is in sentence case (first character upper-, others lowercase)
6. indicator whether the token contains at least one digit
7. indicator of token being a punctuation symbol

These features were evaluated both in isolation and in combination on the trial set. The best performing combination was then used to train the final model.

3.3.3 Feature selection

Evaluating all combinations of the above features, we found that the combination of features 2 and 4 produced the best results on the trial set. For computational efficiency, when selecting the best performing feature subset, we considered lower feature count to overrule a slight increase in performance.

4 Results

Table 2 and Table 3 summarize the results for the statistical and CRF models and their AND and OR combinations on the trial and on the evaluation sets, respectively. For the latter, we used naturally all available labeled data (train and trial sets) for training. Numbers shown correspond to the output of the official evaluation tool. Results on the combination OR represent our official shared task evaluation.

5 Discussion

In the development scenario (Table 2), the main difference between the statistical and CRF model was that the former was superior in recall while the latter in precision. It was thus unclear which of the combinations OR and AND would perform better, we chose OR, the combination method which performed better on the trial set. Unfortunately, the rank of combination methods was different when measured on the evaluation set (Table 3). A possible explanation for this non-extrapolability is the different prior probability of speculative sentences in each set, e.g., 17.9% on the train set while 22.9% on the trial set and 15.8% on the evaluation set.

While using only a minimal amount of features, both of our models were on par with other participants' solutions. Overfitting was observed by the statistical model only (14% drop in precision on the evaluation set), the CRF model showed more consistent behavior across the datasets.

	Model			
	Statistical	CRF	Combination AND	Combination OR
Precision (%)	84.4	92.3	93.9	83.6
Recall (%)	88.6	78.7	75.4	91.8
F-measure (%)	**86.4**	**85.0**	**83.6**	**87.5**

Table 2: Results on trial set (development)

	Model			
	Statistical	CRF	Combination AND	Combination OR
Precision (%)	70.5	87.0	88.0	70.1
Recall (%)	89.4	82.7	81.0	91.0
F-measure (%)	**78.8**	**84.8**	**84.4**	**79.2**

Table 3: Results on evaluation set

6 Conclusion

We presented our method to identify hedging in biomedical literature, and its evaluation at the CoNLL-2010 shared task. We solved the sentence level assertion classification problem by using an ensemble of statistical and CRF models that identify speculation cue phrases. The non-extrapolability of the combination methods' performance observed emphasizes the sensitivity of ensemble methods to the distributions of the datasets they are applied to. While using only a minimal set of standard surface features, our CRF model was on par with participants' systems.

Acknowledgement

D. Tikk was supported by the Alexander-von-Humboldt Foundation.

References

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*, 2001:34–301.

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. ACL.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *BioNLP '09: Proc. of the Workshop on BioNLP*, pages 1–9, Morristown, NJ, USA. ACL.

Andrew K. McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 992–999, Prague, Czech Republic, June. ACL.

Arzucan Özgür and Dragomir R. Radev. 2009. Detecting speculations and their scopes in scientific text. In *EMNLP '09: Proc. of Conf. on Empirical Methods in Natural Language Processing*, pages 1398–1407, Morristown, NJ, USA. ACL.

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. ACL.

Özlem Uzuner, Xiaoran Zhang, and Tawanda Sibanda. 2009. Machine Learning and Rule-based Approaches to Assertion Classification. *Journal of the American Medical Informatics Association*, 16(1):109–115.

Özlem Uzuner. 2008. Second i2b2 workshop on natural language processing challenges for clinical records. In *AMIA Annual Symposium Proceedings*, pages 1252–3.

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.