Bridging the Gap Between Scope-based and Event-based Negation/Speculation Annotations: A Bridge Not Too Far

Pontus Stenetorp[1] **Sampo Pyysalo**[2,3] **Tomoko Ohta**[2,3]
Sophia Ananiadou[2,3] **and Jun'ichi Tsujii**[2,3,4]
[1]Department of Computer Science, University of Tokyo, Tokyo, Japan
[2]School of Computer Science, University of Manchester, Manchester, United Kingdom
[3]National Centre for Text Mining, University of Manchester, Manchester, United Kingdom
[4]Microsoft Research Asia, Beijing, People's Republic of China
{pontus,smp,okap}@is.s.u-tokyo.ac.jp
sophia.ananiadou@manchester.ac.uk
jtsujii@microsoft.com

Abstract

We study two approaches to the marking of extra-propositional aspects of statements in text: the task-independent cue-and-scope representation considered in the CoNLL-2010 Shared Task, and the tagged-event representation applied in several recent event extraction tasks. Building on shared task resources and the analyses from state-of-the-art systems representing the two broad lines of research, we identify specific points of mismatch between the two perspectives and propose ways of addressing them. We demonstrate the feasibility of our approach by constructing a method that uses cue-and-scope analyses together with a small set of features motivated by data analysis to predict event negation and speculation. Evaluation on BioNLP Shared Task 2011 data indicates the method to outperform the negation/speculation components of state-of-the-art event extraction systems.

The system and resources introduced in this work are publicly available for research purposes at: *https://github.com/ninjin/eepura*

1 Introduction

Understanding extra-propositional aspects of texts is key to deeper understanding of statements contained in natural language texts. Extra-propositional aspects such as the polarity of key statements have long been acknowledged to be critical for user-facing applications such as information retrieval (Friedman et al., 1994; Hersh, 1996). In recognition of this need, a number of recent information extraction (IE) resources involving structured representations of text statements have explicitly included some marking of certainty and polarity (LDC, 2005; Kim et al., 2009; Saur and Pustejovsky, 2009; Kim et al., 2011a; Thompson et al., 2011).

Although extra-propositional aspects are recognised as important, there is no clear consensus on how to address their annotation and extraction from text. Some comparatively early efforts focused on the detection of negation *cue* phrases associated with specific (previously detected) terms through regular expression-based rules (Chapman et al., 2001). A number of later efforts identified the scope of negation cues with phrases in constituency analyses in sentence structure (Huang and Lowe, 2007). Drawing in part on this work, the BioScope corpus (Vincze et al., 2008) applied a representation where both cues and their associated *scopes* are marked as contiguous spans of text (Figure 1 bottom). This approach was also applied in the CoNLL-2010 Shared Task (Farkas et al., 2010), in which 13 participating groups proposed approaches for Task 2, which required the identification of uncertainty cues and their associated scopes in text. In the following, we will term this task-independent, linguistically-motivated approach as the *cue-and-scope* representation (please see Vincze et al. (2008) for details regarding the representation).

For IE efforts, more task-oriented representations are commonly applied. In an effort to formalise and drive research for extracting structured representations of statements regarding molecular biology, the ongoing series of BioNLP shared tasks have addressed biomedical Event Extraction (EE) (Kim et al., 2009; Kim et al., 2011a). The extra-propositional targets of negation and speculation

Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM-2012), pages 47–56, Jeju, Republic of Korea, 13 July 2012.

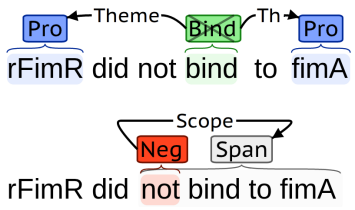

Figure 1: Example illustrating cue-and-scope and event-based negation marking. "Crossing-out" marks events as negated. PRO, TH and NEG are abbreviations for PROTEIN, THEME and NEGATION, respectively.

of extracted events were already included in the first task in the series, using a representation where events can be assigned "flags" to mark them as being negated, speculated, or both (Figure 1 upper). Due to space limitations we refer the reader to Kim et al. (2009) for a detailed explanation of the representation; similar representations have been applied also in previous event extraction tasks (LDC, 2005).

There are a number of ways in which task-oriented, event-based approaches could benefit from the existing linguistically-oriented cue-and-scope methods for identifying extra-propositional aspects of text statements. However, there has been surprisingly little work exploring the combination of the approaches, and comparatively few methods addressing the latter task in detail. Only three out of the 24 participants in the BioNLP Shared Task 2009 submitted results for the non-mandatory negation/speculation task, and although negation and speculation were also considered in three main tasks for the 2011 follow-up event (Kim et al., 2011a), the trend continued, with only two participants addressing the negation/speculation aspects of the task. We are aware of only two studies exploring the relationship between the cue-and-scope and event-based representations: in a manual analysis of scope overlap with tagged events, Vincze et al. (2011) identified a number of issues and mismatches in annotation scope and criteria, which may explain in part the lack of methods combining these two lines of research. Kilicoglu and Bergler (2010) approached the problem from the opposite direction and used an existing EE system to extract cue-and-scope annotations in the CoNLL-2010 Shared Task.

In this work, we take a high-level perspective, seeking to bridge the linguistically oriented framework and the more application-oriented event framework to overcome the mismatches demonstrated by Vincze et al. (2011). Specifically, we aim to determine how cue-and-scope recognition systems can be used to produce a state-of-the-art negation/speculation detection system for the EE task.

2 Resources

Several existing resources can support the investigation of the relationship between the linguistically-oriented and task-oriented perspectives on negation/speculation detection. In this study, we make use of the following resources.

First, we study the three BioNLP 2011 Shared Task corpora that include annotation for negation and speculation: the GE, EPI and ID main task corpora (Table 1). Second, we make use of supporting analyses provided for these corpora in response to a call sent by the BioNLP Shared Task organisers to the developers of third-party systems (Stenetorp et al., 2011). Specifically, we use the output of the BiographTA NeSp Scope Labeler (here referred to as *CLiPS-NESP*) (Morante and Daelemans, 2009; Morante et al., 2010) provided by the University of Antwerp CLiPS center. This system provides cue-and-scope analyses for negation and speculation and was demonstrated to have state-of-the-art performance at the relevant CoNLL-2010 Shared Task. Finally, we make use of the event analyses created by systems that participated in the BioNLP Shared Task, made available to the research community for the majority of the shared task submissions (Pyysalo et al., 2012). These analyses represent the state-of-the-art in event extraction and their capability to detect event structures as well as marking them for negation and speculation.

The above three resources present us with many opportunities to relate scope-based annotations to three highly relevant event-based corpora containing negation/speculation annotations.

3 Manual Analysis

To gain deeper insight into the data and the challenges in combining the cue-and-scope and event-oriented perspectives, we performed a manual analysis of the corpus annotations using the manually

Name	Negated Events		Speculated Events		Negated Spans	Speculated Spans	Publication
EPI	103	(5.6%)	70	(3.8%)	561	1,032	Ohta et al. (2011)
GE	759	(7.4%)	623	(6.0%)	1,308	1,968	Kim et al. (2011b)
ID	69	(3.3%)	26	(1.2%)	415	817	Pyysalo et al. (2011)

Table 1: Corpora used for our experiments along with annotation statistics for their respective training sets. The parenthesised values are the relative proportion of negated/speculated event annotations.

Occ. (Ratio)	EPI	ID
Covered	26 (15.03%)	52 (56.52%)
Not-covered	135 (78.03%)	38 (41.30%)
Error-in-gold	12 (6.94%)	2 (2.18%)
Morphological	48 (27.75%)	11 (11.96%)
Hypothesis	44 (25.43%)	15 (16.30%)
Ellipsis	5 (2.89%)	0 (0.00%)
Argument-only	2 (1.16%)	10 (10.87%)

Table 2: Results from the Manual Data Analysis of the EPI and ID test sets.

created BioNLP Shared Task training data event annotations, and the automatic annotations created for this data by the CLiPS-NESP system. The test data was held out and was not directly examined at any point of our study. We performed the analysis specifically on the EPI and ID corpora, as the GE corpus training set texts overlap with the training data for the CLiPS-NESP system (BioScope corpus), and results on this data would thus not reflect the performance of the system on unseen data, and a comparison of the GE and BioScope gold annotations was previously performed by Vincze et al. (2011).

The analysis was performed by an experienced annotator with a doctoral degree in a related field in biology, who individually examined each of the events marked as negated and speculated in the EPI and ID training corpora. For the analysis, the CLiPS-NESP system output was super-imposed onto the BioNLP Shared Task event annotations.

The annotator was asked to assign three primary flags for each event that was marked as negated or speculated: *Covered* if the event trigger was covered by span(s) of the correct type with a correct cue in the cue-and-span analysis, *Not-covered* if not *Covered*, and *Error-in-gold* if the negation/speculation flag on the event annotation was itself incorrect. We also identified a number of additional properties that initial analysis suggested to frequently characterise instances where the coverage of the cue-and-scope system is lacking: *Morphological* was assigned if the negation/speculation of an event could be inferred only from the morphology of the word expressing the event, rather than from cue words in its context (e.g. *unphosphorylated*, *non-glycosylated*); *Hypothesis* for cases where speculation is marked for events stated as hyphotheses[1] under consideration, e.g. "We analysed the methylation status of MGMT"; *Ellipsis* for cases where the modified expression is elided (e.g. "A was phosphorylated but B was not"); and *Argument-only* if the CLiPS-NESP output had marked the argument of an event as negated rather than the event trigger (we use argument in the sense it is used in the BioNLP Shared Tasks, for example, in Figure 1 upper, the two arguments of the event are "fMimR" and "fimA").

The results of the analysis are summarised in Table 2. We find that that the system shows a clear difference in coverage depending on the dataset. For the ID dataset, a majority of the annotations are covered by the appropriate spans, while only a small minority are covered for EPI. Instead, the EPI dataset contains a significant portion of events where extra-propositional aspects can only be distinguished by the morphology of the word expressing the event (all *Morphological* cases were negation) as well as events marked as speculated due to being expressed as hypotheses under study.

The analysis thus identified specific ways in which the applicability of negation-detection systems using a span-and-scope representation could be improved for some tasks.

[1] While it is arguable whether such cases represent speculation (Vincze et al., 2008), separation from affirmatively made claims is clearly motivated for many applications.

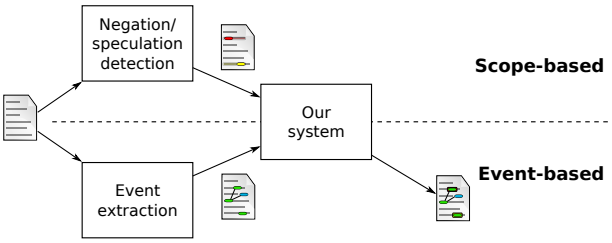

Figure 2: An illustration of our approach.

4 Methods

We next introduce the methods we apply for assigning negation and speculation flags to extracted events.

4.1 Approach

To focus on the extra-propositional aspects of event extraction, we only consider the assignment of the negation and speculation flags, not the extraction of the event structures that these mark. To our knowledge, no previous work studying this subtask in isolation from event extraction exists. Thus, in order to be able to relate the performance of the methods we consider to the performance of previously proposed approaches, it is necessary to base the negation and speculation detection on an event extraction analysis. For this reason, we construct our methods using system outputs for systems participating in the BioNLP Shared Task 2011, in effect creating a negation/speculation processing stage for a pipeline system where the previous stage is the completion of event analysis without negation/speculation detection (Figure 2).

Our methods thus take extracted events as input and attempt to enrich the output with negation and speculation annotations. This enables us to produce a general system with the potential to be applied together with any existing event extraction system. Additionally, this allows us to directly compare our system output with that of the negation/speculation components of previously proposed monolithic systems by removing the existing negation and speculation output from submissions including this and recreating these annotations using our methods.

4.2 Rule-based Methods

The most straightforward way of carrying over information from scope-based to event-based annotations is to consider any event structure for which the word or words stating the event (i.e. the event trigger) is within the scope of a negation or speculation be negated or speculated (respectively). We implemented this simple heuristic as our initial rule-based method.

One relatively common category of cases where this heuristic fails that was identified in analysis relates to events that take other events as arguments. Consider, for example, the case illustrated in Figure 3. The speculation span is correctly identified as covering the statement "FimR modulates mfa1 expression", and the event expressed through "modulates" is identified as speculated. However, the nested event, the expression of *mfa1*, is not speculated. To cover this case, we implemented what we refer to as the *root-heuristic*, which prevents the propagation of negation/speculation marking from scopes to events that are the arguments of another event contained in the same scope. The second rule-based method we consider incorporates this additional heuristic.

Preliminary development set experiments indicated that while the root-heuristic could improve precision, the performance of the rule-based methods remained poor, in particular on the EPI dataset. The results of the manual analysis (Section 3) suggested this to trace in particular to two main issues, namely differences between annotation criteria between BioScope and the shared task data (as noted also by Vincze et al. (2011)) and events which are negated not by external cues but by morphological alternations of the event trigger, such as "unphosphorylated" expressing the absence of phosphorylation. As it would have been difficult to systematically incorporate both morphology and context into the rule-based method without compromising the generality of the approach, we opted to move to a machine learning framework for further method development. This allows us to continue to make use of the existing cue-and-scope annotations while exploring the effects of other aspects of the text and maintaining generality through retraining.

4.3 Machine Learning-based Methods

In developing a machine learning-based approach to the negation/speculation task, we aimed to identify and evaluate a minimal set of features directly mo-

Feature	Example Value(s)
Heuristic	ROOT/NON-ROOT
Heuristic-Cue	possibility
Heuristic-Span	One, possibility, ...
Trigger-Text	non-phosphorylated
Trigger-Prefixes	no, non, non-, ...
Trigger-Preceding-Context	is, that, ...
Trigger-Proceeding-Context	mfa1, expression, ...

Table 3: Machine learning features. The features are categorised into three groups: features based on cue-and-scope based heuristics (top), non-contextual features derived from the event trigger (middle), and features derived from the context of the event trigger (bottom). These three feature sets are abbreviated as E, M and C, respectively.

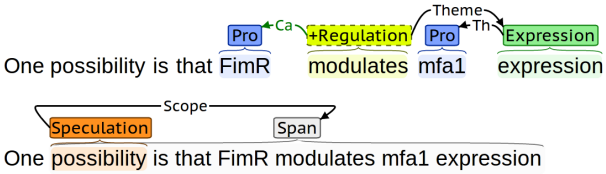

Figure 3: Example of a speculation span containing two events, of which only one is speculated (marked by a dashed border).

tivated by the analysis of the data and to use the cue-and-scope analyses as much as possible. In particular, we wanted to avoid features requiring computationally expensive analyses such as full parsing or replicating the type of analyses performed by the CLiPS-NESP system, focusing rather on specific points where its output does not meet the needs of the event-based approach.

We introduced features representing the heuristics described in Section 4.2, marking each case as being either a root or non-root event in its scope (if any). Drawing further on the cue-and-scope analysis, we included as features the cue word and bag-of-words features for all tokens in the scope (using simple white-space tokenisation). To address the issues identified in manual analysis, we introduced features for the event trigger text as well as character-based prefixes of lengths 2 to 7 of the, intended primarily to capture morphological negation.

All features presented above are derived only from those parts of the sentence already marked either by the event extraction or the cue-and-scope system. However, due to the differences in annotation guidelines for speculation annotations, we expect that the scope-based system will fail to mark a significant portion of the speculation annotations. To allow the system to learn to detect these, we introduce a minimal set of contextual features, limited to a bag-of-words representation of the three words preceding and following the event trigger.

5 Experiments

We perform two sets of experiments, the first to evaluate our approach on gold annotations to give a fair upper-limit to how well our negation/speculation detection system could perform under ideal settings, and the second to enrich the output of an event extraction system with negation and speculation annotations, to evaluate real-world performance and to allow direct comparison of our methods with those incorporated in monolithic event extraction and negation/speculation detection systems.

5.1 Corpora

For our experiments we used the GE, EPI and ID corpora of the BioNLP Shared Task 2011 (Table 1). We note that while the GE training set texts overlap with the BioScope corpus used to train the CLiPS-NESP system, the GE test set does not, and thus test set results are not expected to be overfit.

We noted when performing development set experiments that training machine learning-based methods on the negation/speculation annotations of the event-annotated corpora was problematic due to the sparseness of these flags in the annotation. To address this issue, we merge the training data of the three corpora in all experiments with machine learning methods.

5.2 Baseline methods

We use the event analyses created by the UTurku (Björne and Salakoski, 2011) and UConcordia (Kilicoglu and Bergler, 2011) systems for the BioNLP 2011, the only systems that included negation and speculation analyses. To investigate the impact on a system that did not include a negation/speculation component, we further consider analyses created

Negation (R/P/F)	EPI	GE	ID
H	29.23/31.67/30.40	53.92/52.84/53.38	44.00/31.88/36.97
HR	27.69/32.73/30.00	53.24/71.89/61.18	44.00/37.93/40.74
M	47.69/20.00/28.18	43.00/25.25/31.82	46.00/26.74/33.82
ME	**60.00**/66.10/62.90	58.36/70.08/63.69	54.00/69.23/60.67
MC	40.00/**74.29**/52.00	58.36/76.34/66.15	52.00/61.90/56.52
MCE	58.46/73.08/**64.96**	**61.77/83.03/70.84**	**58.00/70.73/63.74**

Table 4: Results for Negation for our two heuristics and the four combinations of machine learning features.

Speculation (R/P/F)	EPI	GE	ID
H	13.46/6.48/8.75	**33.77**/18.12/23.58	**54.17**/6.50/11.61
HR	11.54/5.66/7.59	32.79/29.45/31.03	**54.17**/7.98/13.90
M	1.92/0.62/0.93	25.65/10.84/15.24	45.83/10.58/17.19
ME	3.85/12.50/5.88	22.08/42.24/29.00	29.17/28.00/28.57
MC	**51.92/52.94/52.43**	27.27/50.30/35.37	37.50/31.03/33.96
MCE	48.08/51.02/49.50	31.82/**53.85/40.00**	33.33/**42.11/37.21**

Table 5: Results for Speculation for our two heuristics and the four combinations of ML features.

by the FAUST system, which achieved the highest performance at two of the three tasks considered (Riedel et al., 2011). The UTurku system is a pipeline ML-based EE system, while the UConcordia system is strictly rule-based. FAUST is an ML-based model combination system incorporating information from the parser-based Stanford system (McClosky et al., 2011) and the jointly-modelled UMass system (Riedel and McCallum, 2011).

We also performed preliminary experiments for the other released submissions to the BioNLP 2011 Shared Task, but due to space limitations focus only on the three above-mentioned systems.

5.3 Evaluation criteria

We use the primary evaluation criteria of the BioNLP 2011 Shared Task (Kim et al., 2011a) to assure comparability, reporting all results using the standard precision, recall and their harmonic mean (F-score).

5.4 Methods

We apply the rule-based simple heuristic method and its root extension (Section 4.2) as well as Support Vector Machines (SVM) trained with the features introduced in Section 4.3. For the SVM, we separately evaluate models based on all permutations of the feature sets introduced in Table 3. In the results tables we abbreviate the feature set names as done in Table 3 and use H for the heuristic method and R for its root extension. As our machine learning component we use LIBLINEAR (Fan et al., 2008) with a L2-regularised L2-loss SVM model. We optimise the SVM regularisation parameter C using 10-fold cross-validation on the training data.

We use the training, development and test set partition provided by the shared task organisers. In line with standard ML methodology the test set was held out during development and was only used when carrying out the final experiments prior to submitting the manuscript.

6 Results and Discussion

Our initial experiments, building on gold event data (Tables 4 and 5), support our manual analysis, showing nearly uniform performance improvement with additional features. First, we find that the root-heuristic gives an improvement over the original heuristic in four out of six cases. To justify our usage of the cue-and-scope based heuristic feature (E) we find that adding it as a feature improves on the M feature set and the MC feature set, showing that even given context, the cue-and-scope perspective is still useful. The only anomaly is for speculation on the EPI dataset, where adding this heuristic feature actually hampers performance, possibly relating to the

Negation (R/P/F)	EPI	GE	ID
UConcordia	16.92/61.11/26.51	18.43/**43.44**/25.88	22.00/23.91/22.92
UConcordia*	**20.00/70.59/31.17**	**20.14**/42.96/**27.42**	**28.00/31.58/29.68**
UTurku	12.31/38.10/18.60	**22.87/48.85/31.15**	**26.00/44.83/32.91**
UTurku*	**43.08/48.28/45.53**	21.16/38.56/27.33	**26.00**/41.94/32.10
FAUST*	29.23/59.38/39.18	21.50/41.18/28.25	**28.00/46.67/35.00**

Table 6: Results of the Negation enrichment experiment.

Speculation (R/P/F)	EPI	GE	ID
UConcordia	**5.77/8.33/6.82**	**21.10/38.46/27.25**	**8.33**/2.00/3.23
UConcordia*	1.92/4.55/2.70	12.99/29.20/17.98	**8.33/2.22/3.51**
UTurku	30.77/**48.48**/37.65	**17.86/32.54/23.06**	**12.50/18.75/15.00**
UTurku*	**46.15**/47.06/**46.60**	11.04/26.56/15.60	8.33/3.33/4.76
FAUST*	36.54/48.72/41.76	10.39/26.50/14.93	**12.50**/12.50/12.50

Table 7: Results of the Speculation enrichment experiment.

(R/P/F)	EPI	ID
UConcordia	**20.83**/42.14/27.88	49.00/40.27/44.21
UConcordia*	**20.83/42.94/28.05**	**49.20/41.78/45.19**
UTurku	52.69/**53.98**/53.33	**37.85/48.62/42.57**
UTurku*	**54.72**/53.86/**54.29**	37.79/47.76/42.19
FAUST	28.88/44.51/35.03	48.03/**65.97**/55.59
FAUST*	**31.64/45.17/37.21**	**49.20**/64.66/**55.88**

Table 8: Overall scores for the EPI and ID data sets.

F	EPI	GE	ID
UConcordia	57.43	60.68	67.28
UTurku	81.31	66.27	55.84
FAUST	74.91	66.14	67.13

Table 9: Estimated F-score upper-bound for an oracle system precision assigning negation/speculation annotations to events predicted by an up-stream EE system.

sparseness of useful annotations due to the differing annotation guidelines, as noted in manual analysis. The numbers from these initial experiments serve as an upper bound when we proceed to our enrichment experiments, as they do not suffer from the possibility of producing false positives negation/speculation annotations for false positive event structures.

In addition to the above in preliminary experiments we also considered two features inspired by findings made by Vincze et al. (2011). A distance-based feature, measuring the distance in tokens between the cue-word and the event trigger, and also trigger suffixes to capture some cases of morphological speculation ("induced" vs. "inducible"). However, we failed to establish any consistent benefits from these features and only for the EPI dataset did the suffix features improve performance.

For the enrichment evaluation, adding negation/speculation flags to the output of event extraction systems (Tables 6 and 7), our results are somewhat more modest. For negation we see an improvement in four out of six cases, and for speculation in two out of six. Despite the fact that a major limitation to our approach are the false positive events that are propagated from the original EE system, we manage to improve the global score for all data sets where a global score is provided by the organisers (Table 8). We improve a full point in F-score for UTurku on EPI, but only sub-percentage for Faust on ID, the latter most likely since ID contains fewer negation and speculation annotations and the global scores are microaverages over all annotations.

As a final analysis we estimate the upper-bound in F-score performance for all three EE systems (Table 9). We do so by assuming that the recall for events marked by negation and speculation is

equal to that of the overall recall of the up-stream EE system and that negation/speculation annotations assigned by an oracle. What we can see is that there is still room for improvement, both for our enrichment approach and for the EE system's internal negation/speculation components, although recall of the EE output is a limiting factor we can expect further efforts towards improving the extra-propositional aspects of the system to yield performance improvements.

7 Conclusions and Future Work

In this study, we have considered two broad lines of research on extra-propositional aspects of key statements in text, one using the task-independent, linguistically-motivated cue-and-scope representation applied in the recent CoNLL-2010 Shared Task, and the other using the task-oriented flagged-event representation applied e.g. in the ACE and BioNLP Shared Task evaluations. We presented a detailed manual analysis exploring points of disagreement and evaluated in detail rule-based and machine learning-based methods joining state-of-the-art systems representing the two approaches.

Our manual analysis identified a number of phenomena that limit the applicability of existing cue-and-scope based systems to the event extraction task, such as negation expressed through morphological change of words expressing events (e.g. *unphosphorylated*). To address these issues, we proposed a combination of heuristics and simple lexical features, carefully selected to address differences in perspective between the cue-and-scope and event-based frameworks and aiming to complement cue-and-scope analyses for creating task-oriented outputs.

To test our approach, we created a method suitable for use as a component of an event extraction pipeline that incorporates information from a previously proposed state-of-the-art cue-and-scope based negation/speculation detection system and a minimal set of features in an SVM-based system that was shown to enhance and in several cases improve upon the output of existing EE systems. Experiments on the BioNLP Shared Task 2011 EPI and ID datasets demonstrated that the combined approach could improve the results of the best-performing systems at the original task in 5 out of 6 cases, outperforming the highest results reported for any system for these two tasks.

There exist several potential targets for future work on improving our introduced system and to join cue-and-scope and event-based approaches. Since none of the existing EE corpora was constructed with the aim to solely cover negation and speculation annotations and taking into account our finding that merging datasets to compensate for data sparseness is beneficial, it might be worth considering other possible corpora or resources and how they can be used for training our machine learning system.

Also, it would be worthwhile to attempt to combine an existing EE system capable of detecting negation/speculation with our proposed method. Combining the two could yield an ensemble, improving upon an already strong system by bridging the differences in perspectives and tapping into the potential benefits of both approaches.

The system and all resources introduced in this work are publicly available for research purposes at: *https://github.com/ninjin/eepura*

Acknowledgements

The authors would like to thank the anonymous reviewers for their many insightful comments and suggestions for improvements.

This work was funded in part by UK Biotechnology and Biological Sciences Research Council (BBSRC) under project Automated Biological Event Extraction from the Literature for Drug Discovery (reference number: BB/G013160/1), by the Ministry of Education, Culture, Sports, Science and Technology of Japan under the Integrated Database Project and by the Swedish Royal Academy of Sciences.

References

Jari Björne and Tapio Salakoski. 2011. Generalizing Biomedical Event Extraction. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, pages 183–191.

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.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

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*, pages 1–12.

Carol Friedman, Philip O. Alderson, John H.M. Austin, James J. Cimino, and Stephen B. Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

William R. Hersh. 1996. *Information retrieval: a health care perspective*. Springer.

Yuang 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):304–311.

Halil Kilicoglu and Sabine Bergler. 2010. A High-Precision Approach to Detecting Hedges and their Scopes. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 70–77.

Halil Kilicoglu and Sabine Bergler. 2011. Adapting a General Semantic Interpretation Approach to Biological Event Extraction. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, pages 173–182.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011a. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, pages 1–6.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011b. Overview of Genia Event Task in BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, pages 7–15.

LDC. 2005. ACE (Automatic Content Extraction) English Annotation Guidelines for Events. Technical report, Linguistic Data Consortium.

David McClosky, Mihai Surdeanu, and Christopher Manning. 2011. Event Extraction as Dependency Parsing for BioNLP 2011. In *Proceedings of BioNLP 2011*, pages 41–45.

Roser Morante 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*, pages 28–36.

Roser Morante, Vincent Van Asch, and Walter Daelemans. 2010. Memory-based resolution of in-sentence scopes of hedge cues. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task*, CoNLL 2010: Shared Task, pages 40–47.

Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. 2011. Overview of the Epigenetics and Post-translational Modifications (EPI) task of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, pages 16–25.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. Overview of the Infectious Diseases (ID) task of BioNLP Shared Task 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, pages 26–35.

Sampo Pyysalo, Pontus Stenetorp, Tomoka Ohta, Jin-Dong Kim, and Sophia Ananiadou. 2012. New Resources and Perspectives for Biomedical Event Extraction. In *Proceedings of BioNLP 2012 Workshop*. to appear.

Sebastian Riedel and Andrew McCallum. 2011. Robust Biomedical Event Extraction with Dual Decomposition and Minimal Domain Adaptation. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, pages 46–50.

Sebastian Riedel, David McClosky, Mihai Surdeanu, Andrew McCallum, and Christopher D. Manning. 2011. Model Combination for Event Extraction in BioNLP 2011. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, pages 51–55.

Roser Saur and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268. 10.1007/s10579-009-9089-9.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. BioNLP Shared Task 2011: Supporting Resources. In *Proceedings of the BioNLP 2011 Workshop Companion Volume for Shared Task*, pages 112–120.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2011. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12(1):393.

Veronika Vincze, Gyorgy Szarvas, Richard Farkas, Gyorgy Mora, and Janos Csirik. 2008. The Bio-

Scope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Veronika Vincze, Gyorgy Szarvas, Gyorgy Mora, Tomoko Ohta, and Richard Farkas. 2011. Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora. *Journal of Biomedical Semantics*, 2(Suppl 5):S8.