Factuality Detection on the Cheap: Inferring Factuality for Increased Precision in Detecting Negated Events

Erik Velldal
University of Oslo
Department of Informatics
erikve@ifi.uio.no

Jonathon Read
University of Oslo
Department of Informatics
jread@ifi.uio.no

Abstract

This paper describes a system for discriminating between factual and non-factual contexts, trained on weakly labeled data by taking advantage of information implicit in annotations of negated events. In addition to evaluating factuality detection in isolation, we also evaluate its impact on a system for event detection. The two components for factuality detection and event detection form part of a system for identifying negative factual events, or counterfacts, with top-ranked results in the *SEM 2012 shared task.

1 Introduction

The First Joint Conference on Lexical and Computational Semantics (*SEM 2012) is hosting a shared task[1] (Morante and Blanco, 2012) on identifying various elements of negation, and one of the subtasks is to identify negated *events*. However, only events occurring in *factual statements* should be labeled. This paper describes pilot experiments on how to train a *factuality classifier* by taking advantage of implicit information on factuality in annotations of negation. In addition to evaluating factuality detection in isolation, we also assess its impact when embedded in a system for *event detection*. The system was ranked first for the *SEM 2012 subtask of identifying negated events, and also formed part of the top-ranked system in the shared task overall (Read et al., 2012). The experiments presented in this paper further improves on these initial results. Note that the system was designed for submission to the closed track of the shared task, which means development is constrained to using the data provided by the task organizers.

The rest of the paper is structured as follows. We start in Section 2 by giving a brief overview of related work and resources. In Section 3 we then present the problem statement in more detail, along with the relevant data sets. This section also discusses the notion of (non-)factuality assumed in the current paper. We then go on to present and evaluate the factuality classifier in Section 4. In Section 5 we move on to describe the event detection task, which is handled by learning a discriminative ranking function over candidate tokens within the negation scope, using features from paths in constituent trees. Both the event ranking function and the factuality classifier are implemented using the Support Vector Machine (SVM) framework. After evaluating the impact of factuality detection on event detection, we finally provide some concluding remarks and discussion of future directions in Section 6.

2 Related Work

Note that the *SEM 2012 shared task singled out three separate subtasks for the problem of recognizing negation, namely the identification of negation *cues*, their in-sentence *scopes* and the negated factual *events*. Most of the systems submitted for the shared task correspondingly implemented a pipeline consisting of three components, one for each subtask. One thing that set the system of Read et al. (2012) apart from other shared task submissions is that it included a *fourth* component; a dedicated

[1] The web site of the 2012 *SEM Shared Task: `http://www.clips.ua.ac.be/sem2012-st-neg/`

Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM-2012), pages 28–36, Jeju, Republic of Korea, 13 July 2012.

classifier for identifying the *factuality* of a given context. It is this latter problem which is the main focus of the current paper, along with its interactions with the task of identifying events.

The field has witnessed a growing body of work dealing with uncertainty and speculative language over the recent years, and in particular so within the domain of biomedical literature. These efforts have been propelled not least by the several shared tasks that have targeted such phenomena. The shared task at the 2010 Conference on Natural Language Learning (CoNLL) focused on speculation detection for the domain of biomedical research literature (Farkas et al., 2010), with data sets based on the BioScope corpus (Vincze et al., 2008) which annotates so-called speculation *cues* along with their *scopes*. The BioNLP shared tasks of 2009 and 2011 mainly concerned recognizing bio-molecular events in text, but optional subtasks involved detecting whether these events were affected by speculation or negation. The data set used for this task is the Genia event corpus (Kim et al., 2008) which annotates the uncertainty of events according to the three labels *certain*, *probable* and *doubtful* (but without explicitly annotating cue words or scope as in BioScope).

The best performer in the BioNLP 2011 supporting task of detecting speculation modification of events, the system of Kilicoglu and Bergler (2011), achieved an end-to-end F_1 of 27.25 using a manually compiled dictionary of trigger expressions together with a set of rules operating on syntactic dependencies for identifying events and event modification. Turning to the task of identifying speculation cues in the BioScope data, current state-of-the-art systems, implementing simple supervised classification approaches on the token- or sequence-level, achieves F_1-scores of well above 80 (Tang et al., 2010; Velldal et al., 2012). For the task of resolving the scopes of these cues, the current best systems obtain end-to-end F_1-scores close to 60 in held-out testing (Morante et al., 2010; Velldal et al., 2012).

Note that the latter reference is from a forthcoming issue of Computational Linguistics specifically on modality and negation (Morante and Sporleder, 2012). In that same issue, Saurí and Pustejovsky (2012) present a linguistically motivated system for factuality profiling with manually crafted rules operating on dependency graphs. Conceptually treating factuality as a perspective that a particular source (speaker) holds toward an event, the system aims to make this attribution explicit. It is developed on the basis of the FactBank corpus (Saurí and Pustejovsky, 2009), containing manual annotations of pairs of events and sources along the dimensions of polarity (*positive*, *negative*, or *underspecified*) and certainty (*certain*, *probable*, *possible*, or *underspecified*.

Prabhakaran et al. (2010) report experiments with *belief tagging*, which in many ways is similar to factuality detection. Their starting point is a corpus of 10.000 words comprising a variety of genres (newswire text, emails, instructions, etc.) annotated for speaker belief of stated propositions (Diab et al., 2009): Propositional heads are tagged as *committed belief* (CB), *non-committed belief* (NCB), or *not applicable* (NA), meaning no belief is expressed by the speaker. To some degree, CB and NCB can be seen as similar to our categories of factuality and non-factuality, respectively. Applying a one-versus-all SVM classifier by 4-fold cross validation, and using wide range of both lexical and syntactical features, Prabhakaran et al. (2010) report F_1-scores of 69.6 for CB, 34.1 for NCB, and 64.5 for NA.

3 Data Sets and the Notion of Factuality

The data we will be using in the current study is taken from a recently released corpus of Conan Doyle (CD) stories annotated for negation (Morante and Daelemans, 2012). The data is annotated with *negation cues*, the in-sentence *scope* of those cues, as well as the negated *event*, if any. The cue is the word(s) or affix indicating a negation, The scope then indicates the maximal extent of that negation, while the event indicates the most basic negated element. In the annotation guidelines, Morante et al. (2011, p. 4) use the term *event* in a rather general sense; “[i]t can be a process, an action, or a state.” The guidelines occasionally also refer to the notion of *negated elements* as encompassing “the main event or property actually negated by the negation cue” (Morante et al., 2011, p. 27). In the remainder of this paper we will simply take *event* to conflate all these senses.

Some examples of annotated sentences are shown below. Throughout the paper we will use angle brackets for marking negation cues, curly brackets for scopes, and underlines for events.

(1) {There was} ⟨no⟩ {answer}.

(2) {I do} ⟨n't⟩ {think that I am a coward} , Watson , but that sound seemed to freeze my very blood .

In the terminology of Saurí and Pustejovsky (2012), the negation cues are negative polarity particles, and all annotated events in the Conan Doyle data will have a negative polarity and thereby represent *counterfacts*, i.e., events with negative factuality. This should not be confused with non-factuality; a counterfactual statement is not uncertain.

Importantly, however, the Conan Doyle negation corpus does not explicitly contain any annotation of factuality. The annotation guidelines specify that "we focus only on annotating information relative to the negative polarity of an event" (Morante et al., 2011, p. 4). However, the guidelines also specify that events should only be annotated for negations that (i) have a scope and that (ii) occur in *factual statements* (Morante et al., 2011, p. 27). (As we only have annotations for the sentence-level it is possible to have a cue without a scope in cases where the cue negates a proposition in a preceding sentence.) The notion of (non-)factuality assumed in the current work will reflect the way it is defined in the Conan Doyle annotation guidelines. Morante et al. (2011) lists the following types of constructions as *not* expressing factual statements (we here show examples from CD^{DEV} for each case):

- Imperatives:

(3) {Do} ⟨n't⟩ {move} , I beg you , Watson .

- Non-factual interrogatives:

(4) {You do} ⟨n't⟩ {believe it} , do you , Watson ?

- Conditional constructions:

(5) If {the law can do} ⟨nothing⟩ we must take the risk ourselves .

- Modal constructions:

(6) {The fault from what I hear may} ⟨not⟩ {have been entirely on one side} .

- Wishes or desires:

(7) " I hope , " said Dr. Mortimer , " that {you do} ⟨not⟩ {look with suspicious eyes upon everyone [. . .]}

- Suppositions or presumptions:

(8) I think , Watson , {a brandy and soda would do him} ⟨no⟩ {harm} .

- Future tense:

(9) {The shadow} has departed and {will} ⟨not⟩ {return} .

Our goal then, will be to correctly identify these cases in order to separate between factual and non-factual contexts before identifying events. Note that, while an event, if present, must always be embedded in the scope, the indicators of factuality are typically found well outside of this scope. The examples also show that non-factuality here encompasses a wider range of phenomena than what is traditionally covered in work on identifying hedging or speculation.

The examples above illustrate how we can take the data to *implicitly* annotate factuality and non-factuality, and we here show how to take advantage of this to train a factuality classifier. For the experiments in this paper we will let positive examples correspond to negations that are annotated with both a scope and an event, while negative examples correspond to scoped negations with no event. For our training and development data (CD^{DEV}; more details below), this strategy gives 738 positive examples and 317 negatives, spread over 930 sentences.

Our weakly labeled data as defined above comes with several limitations of course. The implicit labeling of factuality will be limited to sentences that are negated. We will also not have access to an event in the cases of non-factuality. Neither, do we have any explicit annotation of factuality cue words for these examples. All we have are instances of negation that we know to be within some non-delimited factual or non-factual context. For our experiments here will therefore use the negation cue itself as a place-holder for the abstract notion of context that we are really trying to make predictions about.

Table 1 presents some basic statistics for the relevant data sets. For training and development we will use the negation annotated version of *The Hound of the Baskerville's* (CDH) and *Wisteria Lodge* (CDW) (Morante and Daelemans, 2012). We refer to the combination of these two data sets as CD^{DEV}. For held-out testing we will use the evaluation data sets prepared for the *SEM 2012 shared task; *The Cardboard Box* (CDC) and *The Red Circle* (CDR) (Morante and Blanco, 2012). We will use CD^{EVAL} to refer to the combination of CDC and CDR. Note

			Scoped Negations	
Data set	**Sentences**	**Negations**	**Factual**	**Non-factual**
CDH	3644	984	616	271
CDW	787	173	122	46
CD^{DEV}	4431	1157	738	317
CDC	496	133	87	41
CDR	593	131	86	35
CD^{EVAL}	1089	264	173	76

Table 1: Summary of the Conan Doyle negation data. Note that the total number of negations (column 3) can be smaller than the number of scoped negations (columns 4+5). The reason is that it is possible to have a cue without a scope in cases where the cue negates a proposition in a preceding sentence (which would not be reflected in these sentence-level annotations). The numbers in the column 'Factual' correspond to scoped negations that include an annotated event.

that the column *Factual* correspond to negations with both a scope and event (i.e., positive examples, in terms of factuality classification), while the *Non-factual* column correspond to negations with scope only and no event (negative examples).

4 Factuality Detection

Having described how we abstractly define our training data above, we can now move on to describe our experiments with training a factuality classifier. It is implemented as a linear binary SVM classifier, estimated using the SVM^{light} toolkit (Joachims, 1999). We start by describing the feature types in Section 4.1 and then present results in Section 4.2.

4.1 Features

The feature types we use are mostly variations over bag-of-words (BoW) features. We include left/right oriented BoW features centered on the negation cue, recording forms, lemmas, and PoS, and using both unigrams and bigrams. These features are extracted both from the sentence as a whole, and from a local window of six tokens to each side of the cue. The optimal window size and the order of n-grams was determined empirically.

The reason for including both local and sentence-level BoW features is that we would like to be able to assign different factuality labels to different instances of negation within the same sentence, but at the same time experiments showed sentence-level features to be very important.

Note that, ideally our features should be centered on the negated event, but since this information is only available for factual contexts, we instead take the negation cue as our starting point. In practice, this seems to provide a good approximation, however, given that the negated event is typically found in close vicinity of the negation cue.

In addition to the BoW type features we have features explicitly recording the first full-stop punctuation following the negation cue (i.e., '.', '!', or '?') as well as whether there is an *if* to the left. Note that, although this information is already implicit in the BoW features, the model appeared to benefit from having them explicitly coupled with the cue itself.

We also experimented with several other features that were not included in the final configuration. These included distance to co-occurring verbs, and modal verbs in particular. We also recorded the presence of speculative verbs based on various word lists manually extracted from the training data. None of these features appeared to contribute information not already present in the simple BoW features.

4.2 Results

Table 2 provides results for our factuality classifier using gold cues and gold scopes. In addition, we also include results for a baseline approach that simply considers all cases to be factual, i.e., the majority class. Note that, in this case the precision (of factuality labeling) is identical to the accuracy, which is close to 70% on both the development and held-out set. The recall for the majority-class baseline is of course at 100%, and the corresponding F_1 is approximately 82 on both data sets. In comparison, our classifier achieves an F_1 of 89.92 for the 10-fold cross-validation runs on the development data and 87.10 on the held-out test data. The accuracy is 83.98 and 80.72, respectively. Across both data sets it is clear that the classifier offers substantial improvements over the baseline. We do however, observe a drop in performance particularly with respect to precision when moving to the held-out set.When inspecting the scores for the two individual sections of the held-out set, CDC and CDR, we find that

Data set	Model	Prec	Rec	F_1	Acc
CD^{DEV}	Baseline	69.95	100.00	82.32	69.95
	Classifier	84.51	96.07	89.92	83.98
CD^{EVAL}	Baseline	69.48	100.00	81.99	69.48
	Classifier	80.60	94.74	87.10	80.72

Table 2: Results for factuality detection (using gold negation cues and scopes), reporting 10-fold cross-validation on CD^{DEV} and held-out testing on CD^{EVAL}.

the classifier seems to have more difficulties with the former. Although recall is roughly the same across the two sections (94.25 and 95.24, respectively, which is again fairly close to the 10-fold recall of 96.07), precision suffers a much larger drop on CDC than CDR (78.85 versus 82.47). On the other hand, it is difficult to reliably assess performance on the individual test sets, given the limited amount of data: There are only 128 relevant test cases in CDC and 121 in CDR. However, there also seems to be signs of overfitting, in that an unhealthy number of the training examples end up as support vectors in the final model (close to 70%).

Note that the F_1-scores cited above targets *factuality* as the positive class label. However, given that this is in fact the majority class it might also be instructive to look at F_1-scores targeting *non-factuality*. (In other words, we will use exactly the same classifier predictions, but compute our scores by letting true positives correspond to former true negatives, false positives to former false negatives, and so on, thereby treating non-factuality as the positive class we are trying to predict.) Of course, while all accuracy scores will remain unchanged, the majority-class baseline yields an F_1 of 0 in this case, as there will be no true positives. Table 3 lists the non-factuality scores for the classifier.

Given that we are not aware of any other studies on (non-)factuality detection on this data we are not yet able to directly compare our results against those of other approaches. Nonetheless, we believe the state-of-the-art results cited in Section 2 for related tasks such as belief tagging and identifying speculation cues give reasons for being optimistic about the results obtained with the simple classifier used in these initial pilot experiments.

Data set	Prec	Rec	F_1
CD^{DEV}	77.21	66.25	71.31
CD^{EVAL}	81.25	50.00	61.91

Table 3: Results for non-factuality detection (using gold negation cues and scopes). The scores are based on the same classifier predictions as in Table 2, but treats non-factuality as the positive class.

4.3 Error Analysis and Sample Size Effects

In order to gauge the effect that the size of the training set has on performance we also experimented with leaving out portions of the training examples in our 10-fold cross-validation runs. Figure 1 plots a learning curve showing how classifier performance on CD^{DEV} changes as we incrementally include more training examples. In order to more clearly bring out the contrasts in performance we here plot results against *non*-factuality scores. We also show the size of the training set on a logarithmic scale to better see whether improvements are constant for n-fold increases of data. As can be seen, the learning curve appears to be growing linearly with the increments in larger training samples and it seems safe to assume that the classifier would greatly benefit from

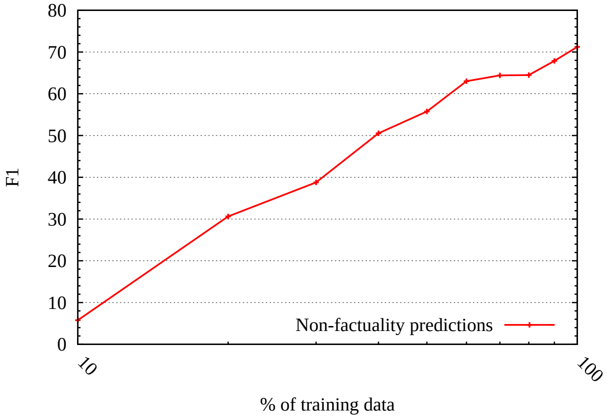

Figure 1: Learning curve showing the effect on F_1 for non-factuality labeling when withdrawing portions of the training partitions (shown on a logarithmic scale) across the 10-fold cross-validation cycles.

additional training data.

This impression is strengthened by a manual inspection of the misclassifications for CD^{DEV}. Quite a number of errors seem related to a combination of scarcity and noise in the data. As a fairly typical example, consider the following negation which the system incorrectly classifies as factual:

(10) “ I presume , sir , ” said he at last , “ that {it was} ⟨not⟩ {merely for the purpose of examining my skull that you have done me the honour to call here last night and again today} ? ”

One could have hoped that the BoW features recording the presence of *presume* would have tipped this prediction toward non-factual. However, while there are ten occurrences of *presume* in CD^{DEV}, only three of these are in contexts that we can actually use as part of our factuality training data. Apart from the one in Example (10), these are shown in (11) and (12) below, both of which indicate factual contexts (given the labeling of an event). We would at least consider Example (11) to reveal an error in the gold annotation here, however.

(11) “ {There is} ⟨no⟩ {other claimant} , I presume ? ”

(12) “ {I presume} ⟨nothing⟩ .

We also get a few errors for incorrectly labeling a context as factual in cases where there are no obvious indicators of non-factuality but the annotation does not mark an event, as in:

(13) “ ⟨Nothing⟩ {of much importance} , Mr. Holmes .

For some of the other errors we observed it would seem that introducing additional features that are sensitive to the syntactic structure could be beneficial. For example, consider sentence (14) below where we incorrectly classify the first negation as non-factual;

(14) [. . .] {I had brought it} only to defend myself if attacked and ⟨not⟩ {to shoot {an} ⟨un⟩{armed man} who was} running {away} .

The error is most likely due to overgeneralizing from the presence of *if*. By letting the lexical features be extracted from a context constrained by the syntax tree rather than a simple sliding window, such errors might be avoided.

For some more optimistic examples, note that the previously listed examples of non-factuality in (3) through (9) were all selected among cases that were correctly predicted by our classifier.

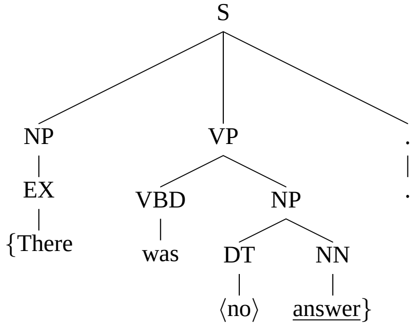

Figure 2: Example of parse tree in the negation data set.

In the next section we move on to describe a system for identifying negated events and assess the impact of the factuality classifier on this task (recall from Section 3 that only negations occurring in factual statements should be assigned an event).

5 Event Detection

To identify events in factual instances of negation[2] we employ an automatically-learned discriminative ranking function. As training data we select all negation scopes that have a single-token[3] event, and generate candidates from each token in the scope. The candidate that matches the event is labeled as correct; all others are labeled as incorrect. For the example sentence in Figure 2 there are three words in the scope and thus three candidates for events: *There*, *was* and *answer*.

5.1 Features

Candidates are primarily described in terms of paths in constituent trees.[4] In particular, we record the full path from a candidate token to the constituent whose projection matches the negation scope (i.e., the most-specific constituent that subsumes all can-

[2] Note that, although one could of course argue that negated events should also be identified for non-factual contexts, that is not how the task is construed in *SEM 2012 shared task or in the Conan Doyle data sets.

[3] To simplify the system we assume that all events are single tokens. It should be noted, however, that 9.85% of events in CD^{DEV} are actually composed of multiple tokens.

[4] Constituent trees from Charniak and Johnson’s Max-Ent reranking parser (2005) were provided by the task organizers.

didates). In Figure 2 this is the `S` root of the tree; the path that describes the correct candidate is `answer/NN/NP/VP/S`. We also record delexicalized paths (e.g., `./NN/NP/VP/S`) and generalized paths (e.g., `./NN//S`), as well as bigrams formed of nodes on the path. Furthermore, we record some surface properties of candidates, namely; lemma, part-of-speech, direction and distance from cue, and position in scope. Finally, we record the lemma and part-of-speech of the token immediately preceding the candidate (development testing showed that information about the token following the candidate was not beneficial).

Based on the features above we learn an SVM-based scoring function using the implementation of ordinal ranking in SVM^{light} (Joachims, 2002). We use a linear kernel and empirically tune the regularization parameter C (governing the trade-off between margin size and errors).

5.2 Results

Similarly to the learning curve shown above for factuality detection, Figure 3 plots the F_1 of event detection on CD$^{\mathrm{DEV}}$ when providing increasing amounts of training data and using gold standard information on factuality. (Note that, except for end-to-end results below, all scores reported in this paper assumes gold negation cues and gold scopes, given that we want to isolate the performance of the event ranker and/or factuality classifier.) We see that the performance is remarkably strong even at 10% of the total data, and increases steadily until around 60%, at which point it appears to be leveling off. It is unclear as to whether or not the ranker would benefit from additional data. We also note differences with respect to the factuality learning curve in Figure 1, both in terms of “entry performance” and overall trend. To some degree, there are general reasons as to why one could expect to see differences in learning curves for a discriminative ranking/regression set-up and a classifier set-up (assuming that the class distribution for the latter is unbalanced, as is typically the case). For a ranker, every item provides useful training data, in the sense that each item provides both positive and negative examples (in our case selected from the candidate tokens within a negation scope). For a classifier, the few items providing examples of the minority class will typically be the most valuable and it will therefore easily be more sensitive to having the training sample restrained. Even so, it seems clear that the factuality detection component and event detection component belong to different ends of the spectrum in terms of sensitivity to sample size.

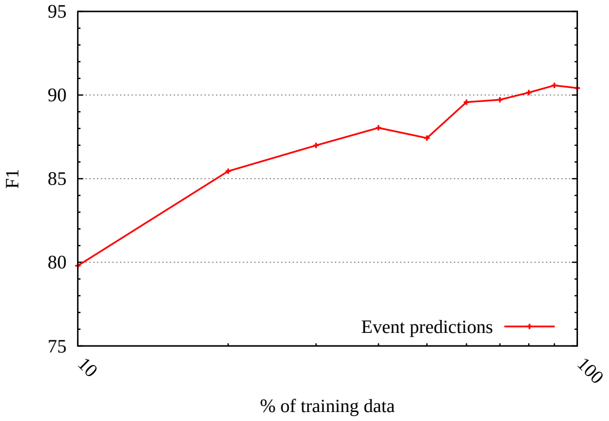

Figure 3: Learning curve showing the effect on F_1 for event detection when using gold factuality and withdrawing portions of the training partitions (shown on a logarithmic scale) across the 10-fold cross-validation cycles.

Table 4 details the results of using the final ranking model to predict negated events. For a comparative baseline, we implemented a basic ranker that uses only the candidate lemma as a single feature. This baseline achieves an F_1 of 73.90 (P=74.01, R=73.80) on CD$^{\mathrm{DEV}}$ when using factuality information inferred from the gold-standard (and testing by 10-fold cross-validation). For comparison, the full ranking model achieves an F_1 of 90.42 (P=90.75, R=90.10) on the same data set, as seen in Table 4.

Of course, the results for event detection using gold-standard factuality also provides the upper bound for what we can achieve using system predicted factuality, i.e., applying the classifier described in Section 4. In order to assess the impact of the factuality classifier we also include results for event detection using the majority-class baseline, which means simply assuming that all instances of negations are factual. Table 4 lists results for event detection using system predicted factuality, compared to results using baseline and gold-standard factuality. We find that the factuality classifier greatly improves precision of the event de-

Data set	**Factuality**	**Prec**	**Rec**	**F_1**
CD^{DEV}	Baseline	62.24	90.10	73.62
	Classifier (10-fold)	78.48	82.98	80.67
	Gold	90.75	90.10	90.42
CD^{EVAL}	Baseline	58.26	84.94	69.11
	Classifier (Held-out)	68.72	80.24	74.03
	Gold	84.94	84.94	84.94

Table 4: Results for event detection using various methods for factuality detection.

tection. As can be expected, however, this comes with a cost in terms of recall. In both 10-fold cross-validation on CD^{DEV} and held-out testing on CD^{EVAL} we find large improvements in F_1, corresponding to error reductions of 26.73% and 15.93% respectively. As expected given the results discussed in Section 4, the improvement is slightly less pronounced for the held-out test results than the 10-fold cross-validated development results. Although the factuality classifier improves substantially over the baseline, it is also clear that a large gap remains toward the "upper bound" results of using gold-standard factuality. We take the results of the pilot experiments described in this paper as a proof-of-concept for using the CD data for training a factuality classifier, and at the same time have high expectations that future experimentation with additional (syntactically oriented) feature types should be able to further advance performance considerably.

Building on the system presented in Velldal et al. (2012), the initial *SEM 2012 shared task submission of Read et al. (2012) also included an SVM negation cue classifier (including support for morphological cues) along with an SVM-based ranking model over syntactic constituents for scope resolution. Coupled with the components for factuality and event detection described above, the end-to-end result for this system on CD^{EVAL} for identifying negated events is F_1=67.02 (P=60.58, R=75.00), making it the top-ranked submission in the shared task.

6 Conclusions and Future Directions

This paper has demonstrated that a classifier for discriminating between factuality and non-factuality can be trained by taking advantage of implicit information on factuality found in the negation annotations of the Conan Doyle corpus (Morante and Daelemans, 2012). Even though the pilot experiments described in this paper use just simple lexical features, the factuality classifier provides substantial improvements over the majority-class baseline. We also present a system for detecting negated events by learning an SVM-based discriminative ranking function over candidate tokens within the negation scope. We show that the factuality classifier proves very useful for improving the precision of event detection. In order to isolate the performance of the event ranker and factuality classifier we have focused on results for gold negation cues and scopes in this paper, although end-to-end results for the full system presented by Read et al. (2012) are also included. The system obtained the best results for identifying negative factual events in the 2012 *SEM shared task.

It is worth noting that there is nothing inherently negation specific about our factuality detection approach *per se*, save for how the training data happens to be extracted in the current study. One reason for using the implicit factuality information in the Conan Doyle negation corpus is the advantage of getting in-domain data, and this also allowed us to stay within the confines of the closed track for the *SEM shared task. For future experiments, however, we would also like to test cross-domain portability by both training and testing the factuality classifier using other annotated data sets such as FactBank, and also add features that incorporate predictions from speculation cue classifiers trained on BioScope.

Acknowledgments

We want to thank Roser Morante and Eduardo Blanco for their effort in organizing the *SEM 2012 shared task and providing the annotations. We also want to thank our colleagues at the University of Oslo (UiO), in particular Lilja Øvrelid and Stephan Oepen who contributed to the shared task submission. Large-scale experimentation was facilitated by the TITAN HPC cluster at UiO. We also thank the anonymous reviewers for their valuable comments and suggestions.

References

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the Forty-Third Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI.

Mona T. Diab, Lori S. Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW 2009)*, pages 68–73, Singapore.

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 14th Conference on Natural Language Learning*, pages 1–12, Uppsala.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods: Support Vector Learning*, pages 41–56. MIT Press, Cambridge, MA.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM International Conference on Knowledge Discovery and Data Mining*, Alberta.

Halil Kilicoglu and Sabine Bergler. 2011. Adapting a general semantic interpretation approach to biological event extraction. In *Proceedings of the BioNLP Shared Task 2011*, pages 173–182, Portland, OR.

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

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, Montreal.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation in Conan Doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):1–38.

Roser Morante, Vincent Van Asch, and Walter Daelemans. 2010. Memory-based resolution of in-sentence scope of hedge cues. In *Proceedings of the 14th Conference on Natural Language Learning*, pages 40–47, Uppsala.

Roser Morante, Sarah Schrauwen, and Walter Daelemans. 2011. Annotation of negation cues and their scope: Guidelines v1.0. Technical report, University of Antwerp. CLIPS: Computational Linguistics & Psycholinguistics technical report series.

Vinodkumar Prabhakaran, Owen Rambow, and Mona T. Diab. 2010. Automatic committed belief tagging. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1014–1022, Beijing.

Jonathon Read, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2012. UiO$_1$: Constituent-based discriminative ranking for negation resolution. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, Montreal. Submission under review.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2).

Buzhou Tang, Xiaolong Wang, Xuan Wang, Bo Yuan, and Shixi Fan. 2010. A cascade method for detecting hedges and their scope in natural language text. In *Proceedings of the 14th Conference on Natural Language Learning*, pages 13–17, Uppsala.

Erik Velldal, Lilja Øvrelid, Jonathon Read, and Stephan Oepen. 2012. Speculation and negation: Rules, rankers and the role of syntax. *Computational Linguistics*, 38(2).

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).