FBK: Exploiting Phrasal and Contextual Clues for Negation Scope Detection

Md. Faisal Mahbub Chowdhury [†‡]
[‡] Fondazione Bruno Kessler (FBK-irst), Trento, Italy
[†] University of Trento, Italy
chowdhury@fbk.eu

Abstract

Automatic detection of negation cues along with their scope and corresponding negated events is an important task that could benefit other natural language processing (NLP) tasks such as extraction of factual information from text, sentiment analysis, etc. This paper presents a system for this task that exploits phrasal and contextual clues apart from various token specific features. The system was developed for the participation in the Task 1 (closed track) of the *SEM 2012 Shared Task (Resolving the Scope and Focus of Negation), where it is ranked 3rd among the participating teams while attaining the highest F_1 score for negation cue detection.

1 Introduction

Negation is a linguistic phenomenon that can alter the meaning of a textual segment. While automatic detection of negation expressions (i.e. cues) in free text has been a subject of research interest for quite some time (e.g. Chapman et al. (2001), Elkin et al. (2005) etc), automatic detection of full scope of negation is a relatively new topic (Morante and Daelemans, 2009; Councill et al., 2010). Detection of negation cues, their scope and corresponding negated events in free text could improve accuracy in other natural language processing (NLP) tasks such as extraction of factual information from text, sentiment analysis, etc (Jia et al., 2009; Councill et al., 2010).

In this paper, we present a system that was developed for the participation in the Scope Detection task of the *SEM 2012 Shared Task[1]. The proposed system exploits phrasal and contextual clues apart from various token specific features. Exploitation of phrasal clues is not new for negation scope detection. But the way we encode this information (i.e. the features for phrasal clues) is novel and differs completely from the previous work (Councill et al., 2010; Morante and Daelemans, 2009). Moreover, the total number of features that we use is also comparatively lower. Furthermore, to the best of our knowledge, automatic negated event/property identification has not been explored prior to the *SEM 2012 Shared Task. So, our proposed approach for this particular sub-task is another contribution of this paper.

The remainder of this paper is organised as follows. First, we describe the scope detection task as well as the accompanying datasets in Section 2. Then in Section 3, we present how we approach the task. Following that, in Section 4, various empirical results and corresponding analyses are discussed. Finally, we summarize our work and discuss how the system can be further improved in Section 5.

2 Task Description: Scope Detection

The Scope Detection task (Task 1) of *SEM 2012 Shared Task deals with intra-sentential (i.e. context is single sentence) negations. According to the guidelines of the task (Morante and Daelemans, 2012; Morante et al., 2011), the scope of a negation cue(s) is composed of all negated concepts and negated event/property, if any. Negation cue(s) is

[1] http://www.clips.ua.ac.be/sem2012-st-neg/

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 340–346,
Montréal, Canada, June 7-8, 2012.

	Training	Development	Test
Total sentence	3644	787	1089
Negation sentences	848	144	235
Negation cues	984	173	264
Cues with scopes	887	168	249
Tokens in scopes	6929	1348	1805
Negated events	616	122	173

Table 1: Various statistics of the training, development and test datasets.

not considered as part of the scope. Cues and scopes may be discontinuous.

The organisers provided three sets of data – training, development and test datasets, all consisting of stories by *Conan Doyle*. The training dataset contains Chapters 1-14 from `The Hound of the Baskervilles`. While development dataset contains `The Adventures of Wisteria Lodge`. For testing, two other stories, `The Adventure of the Red Circle` and `The Adventure of the Cardboard Box`, were released during the evaluation period of the shared task. Table 1 shows various statistics regarding the datasets.

In the training and development data, all occurrences of negation are annotated. For each negation cue, the cue and corresponding scope are marked, as well as the negated event/property, if any. The data is provided in CoNLL-2005 Shared Task format. Table 2 shows an example of annotated data where "un" is the negation cue, "his own conventional appearance" is the scope, and "conventional" is the negated property.

The test data has a format similar to the training data except that only the Columns 1–7 (as shown in Table 2) are provided. Participating systems have to output the remaining column(s).

During a random checking we have found at least 2 missing annotations[2] in the development data. So, there might be few wrong/missing annotations in the other datasets, too.

There were two tracks in the task. For the **closed track**, systems have to be built strictly with information contained in the given training corpus. This includes the automatic annotations that the organizers provide for different levels of analysis (POS tags, lemmas and parse trees). For the **open track**, systems can be developed making use of any kind of external tools and resources.

We participated in the **closed track** of the scope detection task.

[2]Annotations for the following negation cues (and their corresponding scope/negated events) in the development data are missing – {*cue: "no", token no.: 8, sentence no.: 237, chapter: wisteria01*} and {*cue: "never", token no.: 3, sentence no.: 358, chapter: wisteria02*}.

3 Our Approach

We approach the subtasks (i.e. cue, scope and negated event detection) of the Task 1 as sequence identification problems and train three different 1st order Conditional Random Field (CRF) classifiers (i.e. one for each of them) using the MALLET machine learning toolkit (McCallum, 2002). All these classifiers use ONLY the information available inside the training corpus (i.e. training and development datasets) as provided by the task organisers, which is the requirement of the closed track.

3.1 Negation Cue Detection

At first, our system automatically collects a vocabulary of all the positive tokens (i.e. those which are not negation cues) of length greater than 3 characters, after excluding negation cue affixes (if any), from the training data and uses them to extract features that could be useful to identify potential negation cues which are subtokens (e.g. *un*able). We also create a list of highly probable negation expressions (henceforth, *NegExpList*) from the training data based on frequencies. The list consists of the following terms – *nor, neither, without, nobody, none, nothing, never, not, no, nowhere,* and *non*.

Negation cue subtokens are identified if the token itself is predicted as a negation cue by the classifier and has one of the following affixes that are collected from the training data – *less, un, dis, im, in, non, ir*.

Lemmas are converted to lower case inside the feature set. Additional post-processing is done to annotate some obvious negation expressions that are seen inside the training data but sometimes missed by the classifier during prediction on the development data. These expressions include *neither, nobody, save for, save upon,* and *by no means*. A spe-

wisteria01	60	0	Our	Our	PRP$	(S(NP*	_	_	_
wisteria01	60	1	client	client	NN	*)	_	_	_
wisteria01	60	2	looked	look	VBD	(VP*	_	_	_
wisteria01	60	3	down	down	RB	(ADVP*)	_	_	_
wisteria01	60	4	with	with	IN	(PP*	_	_	_
wisteria01	60	5	a	a	DT	(NP(NP*	_	_	_
wisteria01	60	6	rueful	rueful	JJ	*	_	_	_
wisteria01	60	7	face	face	NN	*)	_	_	_
wisteria01	60	8	at	at	IN	(PP*	_	_	_
wisteria01	60	9	his	his	PRP$	(NP*	_	his	_
wisteria01	60	10	own	own	JJ	*	_	own	_
wisteria01	60	11	unconventional	unconventional	JJ	*	un	conventional	conventional
wisteria01	60	12	appearance	appearance	NN	*)))))	_	appearance	_

Table 2: Example of the data provided for *SEM 2012 Shared Task.

Feature name	**Description**
POS_i	Part-of-speech of $token_i$
$Lemma_i$	Lemma form of $token_i$
$Lemma_{i-1}$	Lemma form of $token_{i-1}$
hasNegPrefix	If $token_i$ has a negation prefix and is found inside the automatically created vocabulary
hasNegSuffix	If $token_i$ has a negation suffix and is found inside the automatically created vocabulary
matchesNegExp	If $token_i$ is found in *NegExpList*

Table 3: Feature set for negation cue classifier

cial check is done for the phrase *"none the less"* which is marked as a non-negation expression inside the training data.

Finally, a CRF model is trained using the collected features (see Table 3) and used to predict negation cue on test instance.

3.2 Scope and Negated Event Detection

Once the negation cues are identified, the next tasks are to detect scopes of the cues and negated events which are approached independently using separate classifiers. If a sentence has multiple negation cues, we create separate training/test instance of the sentence for each of the cues.

Tables 4 and 5 show the feature sets that are used to train classifiers. Both the feature sets exclusively use various phrasal clues, e.g. whether the (closest) NP, VP, S or SBAR containing the token under consideration (i.e. $token_i$) and that of the negation cue are different. Further phrasal clues that are exploited include whether the least common phrase of $token_i$ has no other phrase as child, and also list of the counts of different common phrasal categories (starting from the root of the parse tree) that contain $token_i$ and the cue. These latter two types of phrasal clue features are found effective for the negated event detection but not for scope detection.

We also use various token specific features (e.g. lemma, POS, etc) and contextual features (e.g. lemma of the 1st word of the corresponding sentence, position of the token with respect to the cue, presence of conjunction and special characters between $token_i$ and the cue, etc). Finally, new features are created by combining different features of the neighbouring tokens within a certain range of the $token_i$. The range values are selected empirically.

Once scopes and negated events are identified (separately), the prediction output of all the three classifiers are merged to produce the full negation scope.

Initially, a number of features is chosen by doing manual inspection (randomly) of the scopes/negated events in the training data as well analysing syntactic structures of the corresponding sentences. Some of those features (e.g. POS of previous token for scope detection) which are found (empirically) as not useful for performance improvement have been discarded.

Feature name:	Description
Lemma_1	Lemma of the 1st word of the sentence
POS_i	Part-of-speech of token_i
Lemma_i	Lemma of token_i
Lemma_{i-1}	Lemma of token_{i-1}
isCue	If token_i is negation cue
isCueSubToken	If a subtoken of token_i is negation cue
isCcBetCueAndCurTok	If there is a conjunction between token_i and cue
isSpecCharBetCueAndCurTok	If there is a non-alphanumeric token between token_i and cue
Position	Position of token_i : before, after or same w.r.t. the cue
isCueAndCurTokInDiffNP	If token_i and cue belong to different NPs
isCueAndCurTokInDiffVP	If token_i and cue belong to different VPs
isCueAndCurTokInDiffSorSBAR	If token_i and cue belong to different S or SBAR
FeatureConjunctions	New features by combining those of token_{i-2} to token_{i+2}

Table 4: Feature set for negation scope classifier. **Bold** features are the phrasal clue features.

Feature name	Description
Lemma_1	Lemma of the 1st word of the sentence
POS_i	Part-of-speech of token_i
Lemma_i	Lemma of token_i
POS_{i-1}	POS of token_{i-1}
isCue	If token_i is negation cue
isCueSubToken	If a subtoken of token_i is negation cue
isSpecCharBetCueAndCurTok	If there is a non-alphanumeric token between token_i and cue
IsModal	If POS of token_i is MD
IsDT	If POS of token_i is DT
isCueAndCurTokInDiffNP	If token_i and cue belong to different NPs
isCueAndCurTokInDiffVP	If token_i and cue belong to different VPs
isCueAndCurTokInDiffSorSBAR	If token_i and cue belong to different S or SBAR
belongToSamePhrase	If the least common phrase of token_i and cue do not contain other phrase
CPcatBetCueAndCurTok	All common phrase categories (and their counts) that contain token_i and cue
FeatureConjunctions	New features by combining those of token_{i-3} to token_{i+1}

Table 5: Feature set for negated event classifier. **Bold** features are the phrasal clue features.

We left behind two verifications unintentionally which should have been included. One of them is to take into account whether a sentence is a factual statement or a question before negated event detection. The other is to check whether a predicted negated event is found inside the predicted scope of the corresponding negation cue.

4 Results and Discussions

In this section, we discuss various empirical results on the development data and test data. Details regarding the evaluation criteria are described in Morante and Blanco (2012).

4.1 Results on the Development Dataset

Our feature sets are selected after doing a number of experiments by combining various potential feature types. In these experiments, the system is trained on the training data and tested on development data. Due to time limitation we could not do parameter tuning for CRF model training which we assume could further improve the results.

Table 8 shows the results[3] on the development data using the feature sets described in Section 3. There are two noticeable things in these results. Firstly, there is a very high F_1 score (93.29%) obtained for negation cue identification. And secondly, the precision obtained for scope detection (97.92%) is very high as well.

Table 6 shows the results (of negated event iden-

[3]All the results reported in this paper, apart from the ones on test data which are directly obtained from the organisers, reported in this paper are computed using the official evaluation script provided by the organisers.

	TP	FP	FN	Prec.	Rec.	F_1
Using only contextual and token specific features	71	16	46	81.61	60.68	69.61
After adding phrasal clue features	81	17	34	82.65	70.43	76.05

Table 6: Negated event detection results on development data with and without the 5 phrasal clue feature types. The results are obtained using gold annotation of negation cues. Note that, TP+FN is not the same. However, since these results are computed using the official evaluation script, we are not sure why there is this mismatch.

Using negation cues annotated by our system						
	TP	**FP**	**FN**	**Prec.**	**Rec.**	**F_1**
Scope detection	94	2	74	97.92	55.95	71.21
Event detection	63	19	51	76.83	55.26	64.28
Using gold annotations of negation cues						
	TP	**FP**	**FN**	**Prec.**	**Rec.**	**F_1**
Scope detection	103	0	65	100.00	61.31	76.02
Event detection	81	17	34	82.65	70.43	76.05

Table 7: Scope and negated event detection results on development data with and without gold annotations of negation cues. Note that, for negated events, TP+FN is not the same. However, since these results are computed using the official evaluation script, we are not sure why there is this mismatch.

tification) obtained before and after the usage of our proposed 5 phrasal clue feature types (using gold annotation of negation cues). As we can see, there is a significant improvement in recall (almost 10 points) due to the usage of phrasal clues which ultimately leads to a considerable increase (almost 6.5 points) of F_1 score.

4.2 Results on the Official Test Dataset

Table 9 shows official results of our system in the *SEM 2012 Shared Task (closed track) of scope detection, as provided by the organisers. It should be noted that the test dataset is almost 1.5 times bigger than the combined training corpus (i.e. training + development data). Despite this fact, the results of cue and scope detection on the test data are almost similar as those on the development data. However, there is a sharp drop (almost 4 points lower F_1 score) in negated event identification, primarily due to lower precision. This resulted in a lower F_1 score (almost 4.5 points) for full negation identification.

4.3 Further Analyses of the Results and Feature Sets

Our analyses of the empirical results (conducted on the development data) suggest that negation cue identification largely depends on the token itself rather than its surrounding syntactic construction. Although context (i.e. immediate neighbouring tokens) are also important, the significance of a vocabulary of positive tokens (for the identification of negation cue subtokens) and the list of negation cue expressions is quite obvious. In a recently published study, Morante (2010) listed a number of negation cues and argued that their total number are actually not exhaustive. We refrained from using the cues listed in that paper (instead we built a list automatically from the training data) since additional knowledge/resource outside the training data was not allowed for the closed track. But we speculate that usage of such list of expressions as well as an external dictionary of (positive) words can further boost the high performance that we already achieved.

Since scope and negation event detection are dependent on the correct identification of cues, we have done separate evaluation on the development data using the gold cues (instead of predicting the cues first). As the results in Table 7 show, there is a considerable increment in the results for both scope and event detection if the correct annotation of cues are available.

The general trend of errors that we have observed in scope detection is that the more distant a token is from the negation cue in the phrase structure tree (of the corresponding sentence) the harder it becomes for the classifier to predict whether the token should be included in the scope or not. For example, in the sentence "*I am not aware that in my whole life such a thing has ever happened before.*" of the development data, the negation cue "*not*" has scope over the whole sentence. But the scope classifier fails to include the last 4 words in the scope. Perhaps syntactic dependency can provide complementary information in such cases.

As for the negated event identification errors, the majority of the prediction errors (on the development data) occurred for verb and noun tokens which are mostly immediately preceded by the negation cue. Information of syntactic dependency should be

	Gold	System	TP	FP	FN	Prec. (%)	Rec. (%)	F_1 (%)
Cues:	173	156	153	2	20	98.71	88.44	93.29
Scopes (cue match):	168	150	94	2	74	97.92	55.95	71.21
Scopes (no cue match):	168	150	94	2	74	97.92	55.95	71.21
Scope tokens (no cue match):	1348	1132	1024	108	324	90.46	75.96	82.58
Negated (no cue match):	122	90	63	19	51	76.83	55.26	64.28
Full negation:	173	156	67	2	106	97.10	38.73	55.37
Cues B:	173	156	153	2	20	98.08	88.44	93.01
Scopes B (cue match):	168	150	94	2	74	62.67	55.95	59.12
Scopes B (no cue match):	168	150	94	2	74	62.67	55.95	59.12
Negated B (no cue match):	122	90	63	19	51	70.00	55.26	61.76
Full negation B:	173	156	67	2	106	42.95	38.73	40.73
# Sentences: 787		# Negation sentences: 144				# Negation sentences with errors: 97		
% Correct sentences: 87.55						% Correct negation sentences: 32.64		

Table 8: Results on the development data. In the *"B"* variant of the results, *Precision = TP / System*, instead of *Precision = TP / (TP + FP)*.

helpful to reduce such errors, too.

5 Conclusions

In this paper, we presented our approach for negation cue, scope and negated event detection task (*closed track*) of *SEM 2012 Shared Task, where our system ranked 3rd among the participating teams for full negation detection while obtaining the best F_1 score for negation cue detection. Interestingly, according to the results provided by the organisers, our system performs better than all the systems of the *open track* except one (details of these results are described in (Morante and Blanco, 2012)).

The features exploited by our system include phrasal and contextual clues as well as token specific information. Empirical results show that the system achieves very high precision for scope detection. The results also imply that the novel phrasal clue features exploited by our system improve identification of negated events significantly.

We believe the system can be further improved in a number of ways. Firstly, this can be done by incorporating linguistic knowledge as described in Morante (2010). Secondly, we did not take into account whether a sentence is a factual statement or a question before negated event detection. We also did not check whether a predicted negated event is found inside the predicted scope of the corresponding negation cue. These verifications should increase the results more. Finally, previous work reported that usage of syntactic dependency information helps in scope detection (Councill et al., 2010). Hence, this could be another possible direction for improvement.

Acknowledgments

The author would like to thank Alberto Lavelli and the anonymous reviewers for various useful feedback regarding the manuscript.

References

WW Chapman, W Bridewell, P Hanbury, GF Cooper, and BG Buchanan. 2001. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–10.

I Councill, R McDonald, and L Velikovich. 2010. Whats Great and Whats Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59, Uppsala, Sweden.

P Elkin, S Brown, B Bauer, C Husser, W Carruth, L Bergstrom, and D Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making*, 5(1):13.

L Jia, C Yu, and W Meng. 2009. The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness. In

	Gold	System	TP	FP	FN	Prec. (%)	Rec. (%)	F_1 (%)
Cues:	264	263	241	17	23	93.41	91.29	92.34
Scopes (cue match):	249	249	145	18	104	88.96	58.23	70.39
Scopes (no cue match):	249	249	145	18	104	88.96	58.23	70.39
Scope tokens (no cue match):	1805	1825	1488	337	317	81.53	82.44	81.98
Negated (no cue match):	173	154	93	52	71	64.14	56.71	60.20
Full negation:	264	263	96	17	168	84.96	36.36	50.93
Cues B:	264	263	241	17	23	91.63	91.29	91.46
Scopes B (cue match):	249	249	145	18	104	58.23	58.23	58.23
Scopes B (no cue match):	249	249	145	18	104	58.23	58.23	58.23
Negated B (no cue match):	173	154	93	52	71	60.39	56.71	58.49
Full negation B:	264	263	96	17	168	36.50	36.36	36.43
# Sentences: 1089		# Negation sentences: 235			# Negation sentences with errors: 151			
% Correct sentences: 84.94					% Correct negation sentences: 35.74			

Table 9: Results on the *SEM 2012 Shared Task (closed track) test data provided by the organisers. In the *"B"* variant of the results, *Precision = TP / System*, instead of *Precision = TP / (TP + FP)*.

Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), pages 1827–1830, Hong Kong, China.

AK McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu,.

R Morante and E 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 (*SEM 2012)*, Montreal, Canada.

R Morante and W Daelemans. 2009. A Metalearning Approach to Processing the Scope of Negation. In *Proceedings of CoNLL 2009*, pages 28–36, Boulder, Colorado, USA.

R Morante and W Daelemans. 2012. ConanDoyle-neg: Annotation of Negation in Conan Doyle Stories. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

R Morante, S Schrauwen, and W Daelemans. 2011. Annotation of Negation Cues and Their Scope Guidelines v1.0. Technical Report CLiPS Technical Report 3, CLiPS, Antwerp, Belgium.

R Morante. 2010. Descriptive Analysis of Negation Cue in Biomedical Texts. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta.