*SEM 2012 Shared Task: Resolving the Scope and Focus of Negation

Roser Morante
CLiPS - University of Antwerp
Prinsstraat 13, B-2000 Antwerp, Belgium
Roser.Morante@ua.ac.be

Eduardo Blanco
Lymba Corporation
Richardson, TX 75080 USA
eduardo@lymba.com

Abstract

The Joint Conference on Lexical and Computational Semantics (*SEM) each year hosts a shared task on semantic related topics. In its first edition held in 2012, the shared task was dedicated to resolving the scope and focus of negation. This paper presents the specifications, datasets and evaluation criteria of the task. An overview of participating systems is provided and their results are summarized.

1 Introduction

Semantic representation of text has received considerable attention these past years. While early shallow approaches have been proven useful for several natural language processing applications (Wu and Fung, 2009; Surdeanu et al., 2003; Shen and Lapata, 2007), the field is moving towards analyzing and processing complex linguistic phenomena, such as metaphor (Shutova, 2010) or modality and negation (Morante and Sporleder, 2012).

The *SEM 2012 Shared Task is devoted to negation, specifically, to resolving its scope and focus. Negation is a grammatical category that comprises devices used to reverse the truth value of propositions. Broadly speaking, *scope* is the part of the meaning that is negated and *focus* the part of the scope that is most prominently or explicitly negated (Huddleston and Pullum, 2002). Although negation is a very relevant and complex semantic aspect of language, current proposals to annotate meaning either dismiss negation or only treat it in a partial manner.

The interest in automatically processing negation originated in the medical domain (Chapman et al., 2001), since clinical reports and discharge summaries must be reliably interpreted and indexed. The annotation of negation and hedge cues and their scope in the BioScope corpus (Vincze et al., 2008) represented a pioneering effort. This corpus boosted research on scope resolution, especially since it was used in the CoNLL 2010 Shared Task (CoNLL ST 2010) on hedge detection (Farkas et al., 2010). Negation has also been studied in sentiment analysis (Wiegand et al., 2010) as a means to determine the polarity of sentiments and opinions.

Whereas several scope detectors have been developed using BioScope (Morante and Daelemans, 2009; Velldal et al., 2012), there is a lack of corpora and tools to process negation in general domain texts. This is why we have prepared new corpora for scope and focus detection. Scope is annotated in Conan Doyle stories (CD-SCO corpus). For each negation, the cue, its scope and the negated event, if any, are marked as shown in example (1a). Focus is annotated on top of PropBank, which uses the WSJ section of the Penn TreeBank (PB-FOC corpus). Focus annotation is restricted to verbal negations annotated with MNEG in PropBank, and all the words belonging to a semantic role are selected as focus. An annotated example is shown in (1b)[1].

(1) a. [John had] **never** [said as much before]
b. John had never said {as much} before

The rest of this paper is organized as follows. The two proposed tasks are described in Section 2, and the corpora in Section 3. Participating systems and their results are summarized in Section 4. The approaches used by participating systems are described in Section 5, as well as the analysis of results. Finally, Section 6 concludes the paper.

[1] Throughout this paper, negation cues are marked in bold letters, scopes are enclosed in square brackets and negated events are underlined; focus is enclosed in curly brackets.

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 265–274,
Montréal, Canada, June 7-8, 2012.

2 Task description

The *SEM 2012 Shared Task[2] was dedicated to resolving the scope and focus of negation (Task 1 and 2 respectively). Participants were allowed to engage in any combination of tasks and submit at most two runs per task. A pilot task combining scope and focus detection was initially planned, but was cancelled due to lack of participation. We received a total of 14 runs, 12 for scope detection (7 closed, 5 open) and 2 for focus detection (0 closed, 2 open).

Submissions fall into two tracks:

- **Closed track**. Systems are built using exclusively the annotations provided in the training set and are tuned with the development set. Systems that do not use external tools to process the input text or that modify the annotations provided (e.g., simplify parse tree, concatenate lists of POS tags,) fall under this track.
- **Open track**. Systems can make use of any external resource or tool. For example, if a team uses an external semantic parser, named entity recognizer or obtains the lemma for each token by querying external resources, it falls under the open track. The tools used cannot have been developed or tuned using the annotations of the test set.

Regardless of the track, teams were allowed to submit their final results on the test set using a system trained on both the training and development sets. The data format is the same as in several previous CoNLL Shared Tasks (Surdeanu et al., 2008). Sentences are separated by a blank line. Each sentence consists of a sequence of tokens, and a new line is used for each token.

2.1 Task 1: Scope Resolution

Task 1 aimed at resolving the scope of negation cues and detecting negated events. The task is divided into 3 subtasks:

1. Identifying **negation cues**, i.e., words that express negation. Cues can be single words (e.g., *never*), multiwords (e.g., *no longer, by no means*), or affixes (e.g.l *im-*, *-less*). Note that negation cues can be discontinuous, e.g., *neither [. . .] nor*.
2. Resolving the **scope of negation**. This subtask addresses the problem of determining which tokens within a sentence are affected by the negation cue. A scope is a sequence of tokens that can be discontinuous.
3. Identifying the **negated event or property**, if any. The negated event or property is always within the scope of a cue. Only factual events can be negated.

For the sentence in (2), systems have to identify *no* and *nothing* as negation cues, *after his habit he said* and *after mine I asked questions* as scopes, and *said* and *asked* as negated events.

(2) [After his habit he said] **nothing**, and after mine I asked no questions.
After his habit he said nothing, and [after mine I asked] **no** [questions].

2.1.1 Evaluation measures

Previously, scope resolvers have been evaluated at either the token or scope level. The token level evaluation checks whether each token is correctly labeled (inside or outside the scope), while the scope level evaluation checks whether the full scope is correctly labeled. The CoNLL 2010 ST introduced precision and recall at scope level as performance measures and established the following requirements: A true positive (TP) requires an exact match for both the negation cue and the scope. False positives (FP) occur when a system predicts a non-existing scope in gold, or when it incorrectly predicts a scope existing in gold because: (1) the negation cue is correct but the scope is incorrect; (2) the cue is incorrect but the scope is correct; (3) both cue and scope are incorrect. These three scenarios also trigger a false negative (FN). Finally, FN also occur when the gold annotations specify a scope but the system makes no such prediction (Farkas et al., 2010).

As we see it, the CONLL 2010 ST evaluation requirements were somewhat strict because for a scope to be counted as TP, the negation cue had to be correctly identified (strict match) as well as the punctuation tokens within the scope. Additionally, this evaluation penalizes partially correct scopes more than fully missed scopes, since partially correct scopes count as FP and FN, whereas missed scopes count only as FN. This is a standard problem when applying the F measures to the evaluation of sequences. For this shared task we have adopted a slightly different approach based on the following criteria:

- Punctuation tokens are ignored.
- We provide a scope level measure that does not require strict cue match. To count a scope as TP this

[2] www.clips.ua.ac.be/sem2012-st-neg/

measure requires that only one cue token is correctly identified, instead of all cue tokens.
- To count a negated event as TP we do not require correct identification of the cue.
- To evaluate cues, scopes and negated events, partial matches are not counted as FP, only as FN. This is to avoid penalizing partial matches more than missed matches.

The following evaluation measures have been used to evaluate the systems:

- Cue-level F_1-measures (Cue).
- Scope-level F_1-measures that require only partial cue match (Scope NCM).
- Scope-level F_1-measures that require strict cue match (Scope CM). In this case, all tokens of the cue have to be correctly identified.
- F_1-measure over negated events (Negated), computed independently from cues and from scopes.
- Global F_1-measure of negation (Global): the three elements of the negation — cue, scope and negated event — all have to be correctly identified (strict match).
- F_1-measure over scope tokens (Scope tokens). The total of scope tokens in a sentence is the sum of tokens of all scopes. For example, if a sentence has two scopes, one of five tokens and another of seven tokens, then the total of scope tokens is twelve.
- Percentage of correct negation sentences (CNS).

A second version of the measures (Cue/Scope CM/Scope NCM/Negated/Global-B) was calculated and provided to participants, but was not used to rank the systems, because it was introduced in the last period of the development phase following the request of a participant team. In the B version of the measures, precision is not counted as (TP/(TP+FP)), but as (TP / total of system predictions), counting in this way the percentage of perfect matches among all the system predictions. Providing this version of the measures also allowed us to compare the results of the two versions and to check if systems would be ranked in a different position depending on the version.

Even though we believe that relaxing scope evaluation by ignoring punctuation marks and relaxing the strict cue match requirement is a positive feature of our evaluation, we need to explore further in order to define a scope evaluation measure that captures the impact of partial matches in the scores.

2.2 Task 2: Focus Detection

This task tackles focus of negation detection. Both scope and focus are tightly connected. Scope is the part of the meaning that is negated and focus is that part of the scope that is most prominently or explicitly negated (Huddleston and Pullum, 2002). Focus can also be defined as the element of the scope that is intended to be interpreted as false to make the overall negative true.

Detecting focus of negation is useful for retrieving the numerous words that contribute to implicit positive meanings within a negation. Consider the statement *The government didn't release the UFO files {until 2008}*. The focus is *until 2008*, yielding the interpretation *The government released the UFO files, but not until 1998*. Once the focus is resolved, the verb *release*, its AGENT *The government* and its THEME *the UFO files* are positive; only the TEMPORAL information *until 2008* remains negated.

We only target verbal negations and focus is always the full text of a semantic role. Some examples of annotation and their interpretation (Int) using focus detection are provided in (3–5).

(3) Even if that deal isn't {revived}, NBC hopes to find another.
Int: Even if that deal is suppressed, NBC hopes to find another.

(4) A decision isn't expected {until some time next year}.
Int: A decision is expected at some time next year.

(5) ... it told the SEC it couldn't provide financial statements by the end of its first extension "{without unreasonable burden or expense}".
Int: It could provide them by that time with a huge overhead.

2.2.1 Evaluation measures

Task 2 is evaluated using precision, recall and F_1. Submissions are ranked by F_1. For each negation, the predicted focus is considered correct if it is a perfect match with the gold annotations.

3 Data Sets

We have released two datasets, which will be available from the web site of the task: CD-SCO for scope detection and PB-FOC for focus detection. The next two sections introduce the datasets.

WL2	108	0	After	After	IN	(S(S(PP*	_	After	_	_	_	_
WL2	108	1	his	his	PRP$	(NP*	_	his	_	_	_	_
WL2	108	2	habit	habit	NN	*))	_	habit	_	_	_	_
WL2	108	3	he	he	PRP	(NP*)	_	he	_	_	_	_
WL2	108	4	said	say	VBD	(VP*	_	said	said	_	_	_
WL2	108	5	nothing	nothing	NN	(NP*)))	nothing	_	_	_	_	_
WL2	108	6	,	,	,	*	_	_	_	_	_	_
WL2	108	7	and	and	CC	*	_	_	_	_	_	_
WL2	108	8	after	after	IN	(S(PP*	_	_	_	_	after	_
WL2	108	9	mine	mine	NN	(NP*))	_	_	_	_	mine	_
WL2	108	10	I	I	PRP	(NP*)	_	_	_	_	I	_
WL2	108	11	asked	ask	VBD	(VP*	_	_	_	_	asked	asked
WL2	108	12	no	no	DT	(NP*	_	_	_	no	_	_
WL2	108	13	questions	question	NNS	*)))	_	_	_	_	questions	_
WL2	108	14	.	.	.	*)	_	_	_	_	_	_

Figure 1: Example sentence from CD-SCO.

3.1 CD-SCO: Scope Annotation

The corpus for Task 1 is CD-SCO, a corpus of Conan Doyle stories. The training corpus contains *The Hound of the Baskervilles*, the development corpus, *The Adventure of Wisteria Lodge*, and the test corpus *The Adventure of the Red Circle* and *The Adventure of the Cardboard Box*. The original texts are freely available from the Gutenberg Project.[3]

CD-SCO is annotated with negation cues and their scope, as well as the event or property that is negated. The cues are the words that express negation and the scope is the part of a sentence that is affected by the negation cues. The negated event or property is the main event or property actually negated by the negation cue. An event can be a process, an action, or a state.

Figure 1 shows an example sentence. Column 1 contains the name of the file, column 2 the sentence #, column 3 the token #, column 4 the word, column 5 the lemma, column 6 the PoS, column 7 the parse tree information and columns 8 to end the negation information. If a sentence does not contain a negation, column 8 contains "***" and there are no more columns. If it does contain negations, the information for each one is encoded in three columns: negation cue, scope, and negated event respectively.

The annotation of cues and scopes is inspired by the BioScope corpus, but there are several differences. First and foremost, BioScope does not annotate the negated event or property. Another important difference concerns the scope model itself: in CD-SCO, the cue is not considered to be part of the scope. Furthermore, scopes can be discontinuous and all arguments of the negated event are considered to be part of the scope, including the subject, which is kept out of the scope in BioScope. A final difference is that affixal negation is annotated in CD-SCO, as in (6).

(6) [He] declares that he heard cries but [is] **un**[{able} to state from what direction they came].

	Training	**Dev.**	**Test**
# tokens	65,450	13,566	19,216
# sentences	3644	787	1089
# negation sent.	848	144	235
% negation sent.	23.27	18.29	21.57
# cues	984	173	264
# unique cues	30	20	20
# scopes	887	168	249
# negated	616	122	173

Table 1: CD-SCO Corpus statistics.

Statistics for the corpus is presented in Table 1. More information about the annotation guidelines is provided by Morante et al. (2011) and Morante and Daelemans (2012), including inter-annotator agreement.

The corpus was preprocessed at the University of Oslo. Tokenization was obtained by the PTB-compliant tokenizer that is part of the LinGO English Resource Grammar. [4]

[3] http://www.gutenberg.org/browse/authors/d\#a37238

[4] http://moin.delph-in.net/

Apart from the gold annotations, the corpus was provided to participants with additional annotations:

- Lemmatization using the GENIA tagger (Tsuruoka and Tsujii, 2005), version 3.0.1, with the '-nt' command line option. GENIA PoS tags are complemented with TnT PoS tags for increased compatibility with the original PTB.
- Parsing with the Charniak and Johnson (2005) reranking parser.[5] For compatibility with PTB conventions, the top-level nodes in parse trees ('S1'), were removed. The conversion of PTB-style syntax trees into CoNLL-style format was performed using the CoNLL 2005 Shared Task software.[6]

3.2 PB-FOC: Focus Annotation

We have adapted the only previous annotation effort targeting focus of negation for PB-FOC (Blanco and Moldovan, 2011). This corpus provides focus annotation on top of PropBank. It targets exclusively verbal negations marked with MNEG in PropBank and selects as focus the semantic role containing the most likely focus. The motivation behind their approach, annotation guidelines and examples can be found in the aforementioned paper.

We gathered all negations from sections 02–21, 23 and 24 and discarded negations for which the focus or PropBank annotations were not sound, leaving 3,544 instances.[7] For each verbal negation, PB-FOC provides the current sentence, and the previous and next sentences as context. For each sentence, along with the gold focus annotations, PB-FOC contains the following additional annotations:

- Token number;
- POS tags using the Brill tagger (Brill, 1992);
- Named Entities using the Stanford named entity recognizer recognizer (Finkel et al., 2005);
- Chunks using the chunker by Phan (2006);
- Syntactic tree using the Charniak parser (Charniak, 2000);
- Dependency tree derived from the syntactic tree (de Marneffe et al., 2006);
- Semantic roles using the labeler described by (Punyakanok et al., 2008); and
- Verbal negation, indicates with 'N' if that token correspond to a verbal negation for which focus must be predicted.

		Train	Devel	Test
	1 role	2,210	515	672
	2 roles	89	15	38
	3 roles	3	0	2
	All	2,302	530	712
Semantic roles focus belongs to	A1	980	222	309
	AM-NEG	592	138	172
	AM-TMP	161	35	46
	AM-MNR	127	27	38
	A2	112	28	36
	A0	94	23	31
	None	88	19	35
	AM-ADV	78	23	26
	C-A1	46	6	16
	AM-PNC	33	8	12
	AM-LOC	25	4	10
	A4	11	2	5
	R-A1	10	2	2
	Other	40	8	16

Table 2: Basic numeric analysis for PB-FOC. The first 4 rows indicate the number of unique roles each negation belongs to, the rest indicate the counts for each role.

Figure 2 provides a sample of PB-FOC. Knowing that the original focus annotations were done on top of PropBank and that focus corresponds to a single role, semantic role information is key to predict the focus. In Table 2, we show some basic numeric analysis regarding focus annotation and the automatically obtained semantic role labels. Most instances of focus belong to a single role in the three splits and the most common role focus belongs to is A1, followed by AM-NEG, M-TMP and M-MNR. Note that some instances have at least one word that does not belong to any role (88 in training, 19 in development and 35 in test).

4 Submissions and results

A total of 14 runs were submitted: 12 for scope detection and 2 for focus detection. The unbalanced number of submissions might be due to the fact that both tasks are relatively new and the tight timeline (six weeks) under which systems were developed.

ErgTokenization, http://moin.delph-in.net/ReppTop

[5] November 2009 release available from Brown University.

[6] http://www.lsi.upc.edu/~srlconll/srlconll-1.1.tgz

[7] The original focus annotation targeted the 3,993 negations marked with MNEG in the whole PropBank.

Marketers	1	NNS	O	B-NP	(S1(S(NP*)	2	nsubj	(A0*)	*	-	*
believe	2	VBP	O	B-VP	(VP*	0	root	(V*)	*	-	*
most	3	RBS	O	B-NP	(SBAR(S(NP*	4	amod	(A1*	(A0*	-	FOCUS
Americans	4	NNPS	O	I-NP	*)	7	nsubj	*	*)	-	FOCUS
wo	5	MD	O	B-VP	(VP*	7	aux	*	(AM-MOD*)	-	*
n't	6	RB	O	I-VP	*	7	neg	*	(AM-NEG*)	-	*
make	7	VB	O	I-VP	(VP*	2	ccomp	*	(V*)	N	*
the	8	DT	O	B-NP	(NP*	10	det	*	(A1*	-	*
convenience	9	NN	O	I-NP	*	10	nn	*	*	-	*
trade-off	10	NN	O	I-NP	*))))))	7	dobj	*)	*)	-	*
...	11	:	O	O	*	2	punct	*	*	-	*
.	12	.	O	O	*))	2	punct	*	*	-	*

Figure 2: Example sentence from PB-FOC.

	Team	Prec.	Rec.	F1
Open	UConcordia, run 1	60.00	56.88	58.40
	UConcordia, run 2	59.85	56.74	58.26

Table 3: Official results for Task 2.

Some participants showed interest in the second task and expressed that they did not participate because of lack of time. In this section, we present the results for each task.

4.1 Task 1

Six teams (UiO1, UiO2, FBK, UWashington, UMichigan, UABCoRAL) submitted results for the closed track with a total of seven runs, and four teams (UiO2, UGroningen, UCM-1, UCM-2) submitted results for the open track with a total of five runs. The evaluation results are provided in Table 4, which contains the official results, and Table 5, which contains the results for evaluation measures B.

The best Global score in the closed track was obtained by UiO1 (57.63 F_1). The best score for Cues was obtained by FBK (92.34 F_1), for Scopes CM by UiO2 (73.39 F_1), for Scopes NCM by UWashington (72.40 F_1), and for Negated by UiO1 (67.02 F_1). The best Global score in the open track was obtained by UiO2 (54.82 F_1), as well as the best scores for Cues (91.31 F_1), Scopes CM (72.39 F_1), Scopes NCM (72.39 F_1), and Negated (61.79 F_1).

4.2 Task 2

Only one team participated in Task 2, UConcordia from CLaC Lab at Concordia University. They submitted two runs and the official results are summarized in Table 3. Their best run scored 58.40 F_1.

5 Approaches and analysis

In this section we summarize the methodologies applied by participants to solve the tasks and we analyze the results.

5.1 Task 1

To solve Task 1 most teams develop a three module pipeline with a module per subtask. Scope resolution and negated event detection are independent of each other and both depend on cue detection. An exception is the UiO1 system, which incorporates a module for factuality detection. Most systems apply machine learning algorithms, either Conditional Random Fields (CRFs) or Support Vector Machines (SVMs), while less systems implement a rule-based approach. Syntax information is widely employed, either in the form of rules or incorporated in the learning model. Multi-word and affixal negation cues receive a special treatment in most cases, and scopes are generally postprocessed.

The systems that participate in the closed track are machine learning based. The UiO1 system is an adaptation of another system (Velldal et al., 2012), which combines SVM cue classification with SVM-based ranking of syntactic constituents for scope resolution. The approach is extended to identify negated events by first classifying negations as factual or non-factual, and then applying an SVM ranker over candidate events. The original treatment of factuality in this system results in the highest score for both the negated event subtask and the global task.

The UiO2 system combines SVM cue classification with CRF-based sequence labeling. An original aspect of the UiO2 approach is the model represen-

Official results for Task 1

		Cues			Scopes CM			Scopes NCM			Scope Tokens			Negated			Global			% CNS
		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	
Closed track	**UiO1 r2**	89.17	93.56	91.31	83.89	60.64	70.39	83.89	60.64	70.39	75.87	90.08	82.37	60.58	75.00	**67.02**	79.87	45.08	**57.63**	**43.83**
	UiO1 r1	91.42	92.80	92.10	87.43	61.45	72.17	87.43	61.45	72.17	81.99	88.81	**85.26**	60.50	72.89	66.12	83.45	43.94	57.57	42.13
	UiO2	89.17	93.56	91.31	85.71	62.65	**72.39**	85.71	62.65	72.39	86.03	81.55	83.73	68.18	52.63	59.40	78.26	40.91	53.73	40.00
	FBK	93.41	91.29	**92.34**	88.96	58.23	70.39	88.96	58.23	70.39	81.53	82.44	81.98	64.14	56.71	60.20	84.96	36.36	50.93	35.74
	UWashington	88.04	92.05	90.00	82.72	63.45	71.81	82.90	64.26	**72.40**	83.26	83.77	83.51	58.04	50.92	54.25	74.02	35.61	48.09	34.04
	UMichigan	94.31	87.88	90.98	90.00	50.60	64.78	90.00	50.60	64.78	84.85	80.66	82.70	50.00	52.24	51.10	84.27	28.41	42.49	27.23
	UABCoRAL	85.93	85.61	85.77	79.04	53.01	63.46	79.53	54.62	64.76	85.37	68.86	76.23	65.00	38.46	48.33	66.36	27.65	39.04	26.81
Open track	**UiO2**	89.17	93.56	**91.31**	85.71	62.65	**72.39**	85.71	62.65	**72.39**	82.25	82.16	**82.20**	66.90	57.40	**61.79**	78.72	42.05	**54.82**	**41.28**
	UGroningen r2	88.89	84.85	86.82	76.12	40.96	53.26	76.12	40.96	53.26	69.20	82.27	75.17	56.63	65.29	60.65	72.00	27.27	39.56	27.23
	UCM-1	89.26	91.29	90.26	82.86	46.59	59.64	82.86	46.59	59.64	85.37	68.53	76.03	66.67	12.72	21.36	66.28	21.59	32.57	18.72
	UCM-2	81.34	64.39	71.88	67.13	38.55	48.98	66.90	38.96	49.24	58.30	67.70	62.65	46.15	21.18	29.03	42.65	10.98	17.46	11.91
	UGroningen r1	86.90	82.95	84.88	46.38	12.85	20.12	46.38	12.85	20.12	69.69	70.30	69.99	53.94	52.05	52.98	37.74	7.58	12.62	7.66

Table 4: Official results. “r1” stands for run 1 nd “r2” for run 2. CNS stands for Correct Negation Sentences. “CM” stands for Cue Match and “NCM” stands for No Cue Match.

		Cues B			Scopes B CM			Scopes B NCM			Negated B			Global B		
		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Closed track	**UiO1 r2**	86.97	93.56	90.14	56.55	60.64	58.52	56.55	60.64	58.52	58.60	75.00	**65.79**	41.90	45.08	**43.43**
	UiO1 r1	89.09	92.80	90.91	59.30	61.45	60.36	59.30	61.45	60.36	57.62	72.89	64.36	42.18	43.94	43.04
	UiO2	86.97	93.56	90.14	59.32	62.65	60.94	59.32	62.65	60.94	67.16	52.63	59.01	38.03	40.91	39.42
	FBK	91.63	91.29	**91.46**	58.23	58.23	58.23	58.23	58.23	58.23	60.39	56.71	58.49	38.03	40.91	39.42
	UWashington	85.26	92.05	88.52	58.52	63.45	**60.89**	59.26	64.26	**61.66**	53.90	50.92	52.37	32.98	35.61	34.24
	UMichigan	92.80	87.88	90.27	55.51	50.60	52.94	55.51	50.60	52.94	38.25	52.24	44.16	30.00	28.41	29.18
	UABCoRAL	79.58	85.61	82.48	55.23	53.01	54.10	56.90	54.62	55.74	62.50	38.46	47.62	25.70	27.65	26.64
Open track	**UiO2**	86.97	93.56	**90.14**	59.54	62.65	**61.06**	59.54	62.65	**61.06**	63.82	57.40	**60.44**	39.08	42.05	**40.51**
	UGroningen r2	85.82	84.85	85.33	39.84	40.96	40.39	39.84	40.96	40.39	55.22	65.29	59.83	27.59	27.27	27.43
	UCM-1	86.69	91.29	88.93	45.67	46.59	46.13	45.67	46.59	46.13	66.67	12.72	21.36	20.50	21.59	21.03
	UCM-2	72.34	64.39	68.13	41.20	38.55	39.83	41.63	38.96	40.25	44.44	21.18	28.69	12.34	10.98	11.62
	UGroningen r1	83.91	82.95	83.43	12.26	12.85	12.55	12.26	12.85	12.55	52.66	52.05	52.35	7.66	7.58	7.62

Table 5: Results withevaluation measures B. Precision is calculated as: true positives / total of system predictions. “r1” stands for run 1 nd “r2” for run 2. “CM” stands for Cue Match and “NCM” stands for No Cue Match.

Participating institutions:
UiO: University of Oslo; FBK: Fondazione Bruno Kessler & University of Trento; UWashington: University of Washington; UMichigan: University of Michigan; UABCoRAL: CoRAL Lab University of Alabama; UGroningen: University of Groningen; UCM: Complutense University of Madrid.

tation for scopes and negated events, where tokens are assigned a set of labels that attempts to describe their behavior within the mechanics of negation. After unseen sequences are labeled, in-scope and negated tokens are assigned to their respective cues using simple post-processing heuristics.

The FBK system consists of three different CRF classifiers, as well as the UMichigan. A characteristic of the cue model of the UMichigan system is that tokens are assigned five labels in order to represent the different types of negation. Similarly, the UWashington system has a CRF sequence tagger for scope and negated event detection, while the cue detector learns regular expression matching rules from the training set. The UABCoRAL system follows the same strategy, but instead of CRFs it employs SVM Light.

The resources utilized by participants in the open track are diverse. UiO2 reparsed the data with MaltParser in order to obtain dependency graphs. For the rest, the system is the same as in the closed track. The global results obtained by this system in the closed track are higher than the results obtained in the open track, which is mostly due to a higher performance of the scope resolution module. This is the only machine learning system in the open track and the highest performing one.

The UGroningen system is based on tools that produce complex semantic representations. The system employs the C&C tools[8] for parsing and Boxer[9] to produce semantic representations in the form of Discourse Representation Structures (DRSs). For cue detection, the DRSs are converted to flat, non-recursive structures, called Discourse Representation Graphs (DRGs). These DRGs allow for cue detection by means of labelled tuples. Scope detection is done by gathering the tokens that occur within the scope of the negated DRSs. For negated event detection, a basic algorithm takes the detected scope and returns the negated event based on information from the syntax tree within the scope.

UCM-1 and UCM-2 are rule-based systems that rely heavily on information from the syntax tree. The UCM-1 system was initially designed for processing opinionated texts. It applies a dictionary approach to cue detection, with the detection of affixal cues being performed using WordNet. Non-affixal cue detection is performed by consulting a predefined list of cues. It then uses information from the syntax tree in order to get a first approximation to the scope, which is later refined using a set of post-processing rules. In the case of the UCM-2 system an algorithm detects negation cues and their scope by traversing Minipar dependency structures. Finally, the scope is refined with post-processing rules that take into account the information provided by the first algorithm and linguistic clause boundaries.

If we compare tracks, the Global best results obtained in the closed track (57.63 F_1) are higher than the Global best results obtained in the open track (54.82 F_1). If we compare approaches, the best results in the two tracks are obtained with machine learning-based systems. The rule-based systems participating in the open track clearly score lower (39.56 F_1 the best) than the machine learning-based system (54.82 F_1).

Regarding subtasks, systems achieve higher results in the cue detection task (92.34 F_1 the best) and lower results in the scope resolution (72.40 F_1 the best) and negated event detection (67.02 F_1 the best) tasks. This is not surprising, not only because of the error propagation effect, but also because the set of negation cues is closed and comprises mostly single tokens, whereas scope sequences are longer. The best results in cue detection are obtained by the FBK system that uses CRFs and applies a special procedure to detect the negation cues that are subtokens. The best scores for scope resolution (72.40, 72.39 F_1) are obtained by two machine learning components. UWashington uses CRFs with features derived from the syntax tree. UiO2 uses CRFs models with syntactic and lexical features for scopes, together with a set of labels aimed at capturing the behavior of certain tokens within the mechanics of negation. The best scores for negated events (67.02 F_1) are obtained by the UiO1 system that first classifies negations as factual or non-factual, and then applies an SVM ranker over candidate events.

Finally, we would like to draw the attention to the different scores obtained depending on the evaluation measure used. When scope resolution is evaluated with the Scope (NCM, CM) measure, results

[8] http://svn.ask.it.usyd.edu.au/trac/candc/wiki/Documentation

[9] http://svn.ask.it.usyd.edu.au/trac/candc/wiki/boxer

are much lower than when using the Scope Tokens measure, which does not reflect the ability of systems to deal with sequences. Another observation is related to the difference in precision scores between the two versions of the evaluation measures. Whereas for Cues and Negated the differences are not so big because most cues and negated events span over a single token, for Scopes they are. The best Scope NCM precision score is 90.00 %, whereas the best Scope NCM B precision score is 59.54 %. This shows that the scores can change considerably depending on how partial matches are counted (as FP and FN, or only as FN). As a final remark it is worth noting that the ranking of systems does not change when using the B measures.

5.2 Task 2

UConcordia submitted two runs in the open track. Both of them follow the same three component approach. First, negation cues are detected. Second, the scope of negation is extracted based on dependency relations and heuristics defined by Kilicoglu and Bergler (2011). Third, the focus of negation is determined within the elements belonging to the scope following three heuristics.

6 Conclusions

In this paper we presented the description of the first *SEM Shared Task on Resolving the Scope and Focus of Negation, which consisted of two different tasks related to different aspects of negation: Task 1 on resolving the scope of negation, and Task 2 on detecting the focus of negation. Task 1 was divided into three subtasks: identifying negation cues, resolving their scope, and identifying the negated event. Two new datasets have been produced for this Shared Task: the CD-SCO corpus of Conan Doyle stories annotated with scopes, and the PB-FOC corpus, which provides focus annotation on top of PropBank. New evaluation software was also developed for this task. The datasets and the evaluation software will be available on the web site of the Shared Task. As far as we know, this is the first task that focuses on resolving the focus and scope of negation.

A total of 14 runs were submitted, 12 for scope detection and 2 for focus detection. Of these, four runs are from systems that take a rule-based approach, two runs from hybrid systems, and the rest from systems that take a machine learning approach using SVMs or CRFs. Most participants designed a three component architecture.

For a future edition of the shared task we would like to unify the annotation schemes of the two corpora, namely the annotation of focus in PB-FOC and negated events in CD-SCO. The annotation of more data with both scope and focus would allow us to study the two aspects jointly. We would also like to provide better evaluation measures for scope resolution. Currently, scopes are evaluated in terms of F_1, which demands a division of errors into the categories TP/FP/TN/FN borrowed from the evaluation of information retrieval systems. These categories are not completely appropriate to be assigned to sequence tasks, such as scope resolution.

Acknowledgements

We are very grateful to Vivek Srikumar for preprocessing the PB-FOC corpus with the Illinois semantic role labeler, and to Stephan Oepen for preprocessing the CD-SCO corpus. We also thank the *SEM organisers and the ST participants. Roser Morante's research was funded by the University of Antwerp (GOA project BIOGRAPH).

References

Eduardo Blanco and Dan Moldovan. 2011. Semantic Representation of Negation Using Focus Detection. In *Proceedings of the 49th Annual Meeting of the Association for C omputational Linguistics: Human Language Technologies*, pages 581–589, Portland, Oregon, USA. Association for Computational Linguistics.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*, ANLC '92, pages 152–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

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. *J Biomed Inform*, 34:301–310.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 173–180, Ann Arbor.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, NAACL 2000, pages 132–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the IEEE / ACL 2006 Workshop on Spoken Language Technology*. The Stanford Natural Language Processing Group.

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, Uppsala, Sweden. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rodney D. Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Halil Kilicoglu and Sabine Bergler. 2011. Effective bio-event extraction using trigger words and syntactic dependencies. *Computational Intelligence*, 27(4):583–609.

Roser Morante and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proceedings of the 13th Conference on Natural Language Learning*, pages 21–29, Boulder, CO.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories. In *Proceedings of LREC 2012*, Istambul.

Roser Morante and Caroline Sporleder. 2012. Special issue on modality and negation: An introduction. *Computational Linguistics*.

Roser Morante, Sarah Schrauwen, and Walter Daelemans. 2011. Annotation of negation cues and their scope. guidelines v1.0. Technical Report Series CTR-003, CLiPS, University of Antwerp, Antwerp, April.

Xuan-Hieu Phan. 2006. Crfchunker: Crf english phrase chunker.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287, June.

Dan Shen and Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EM NLP-CoNLL)*, pages 12–21.

Ekaterina Shutova. 2010. Models of Metaphor in NLP. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 688–697, Uppsala, Sweden. ACL.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15, Sapporo, Japan. Association for Computational Linguistics.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll-2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning*, page 159177, Manchester.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 467–474, Vancouver.

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

Veronika Vincze, Gyorgy Szarvas, Richard Farkas, Gyorgy Mora, and Janos Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9+.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68, Uppsala, Sweden. University of Antwerp.

Dekai Wu and Pascale Fung. 2009. Semantic Roles for SMT: A Hybrid Two-Pass Model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 13–16, Boulder, Colorado. Association for Computational Linguistics.