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

Roser Morante[1] and Eduardo Blanco[2]

[1]CLiPS - University of Antwerp, [2]Lymba Corporation

Conference, Montreal, June 9, 2012

Outline

Outline

First *SEM Shared Task

- The first *SEM Shared Task combined two tasks related to two aspects of negation:
  - Scope resolution
  - Focus detection
- **Scope** is the part of the meaning that is negated.
- **Focus** is the part of the scope that is most prominently or explicitly negated.

Example

1. [John had never said] [{as much} before]

First *SEM Shared Task

- Two **subtasks**:
  - Task 1: Scope resolution
  - Task 2: Focus detection
- A pilot task combining scope and focus detection was cancelled.
- **Submissions**: a total of 14 runs.
  - 12 for scope detection (7 closed, 5 open)
  - 2 for focus detection (0 closed, 2 open)

Tracks

- **Closed track**
  - Systems are built using exclusively the annotations provided in the training set and are tuned with the development set.
  - Systems do not use external tools to process the input text or modify the annotations provided.
- **Open track**
  - Systems can make use of any external resource or tool.
  - The tools used cannot have been developed or tuned using the annotations of the test set.

Datasets

- **Datasets** available from the web site of the task:
  - **CD-SCO** for scope detection (Morante and Daelemans 2012)
  - **PB-FOC** for focus detection (Blanco and Moldovan 2011)
- **Data format**: column format as in CoNLL Shared Tasks.

Outline

Task description

Resolving the scope of negation cues and detecting negated events

Subtasks

1. Identifying **negation cues**, i.e., words that express negation.
   Single words (never), multiwords (no longer, by no means), affixes (im-, -less), discontinuous (neither [...] nor).
2. Resolving the **scope of negation**. 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.

Task description

Examples

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

CD-SCO corpus

- Conan Doyle stories freely available from the Gutenberg Project.
  - Training: *The Hound of the Baskervilles*.
  - Dev: *The Adventure of Wisteria Lodge*.
  - Test: *The Adventure of the Red Circle* and *The Adventure of the Cardboard Box*.
- Additional annotations provided:
  - Lemmatization using the GENIA tagger.
  - Parsing with the Charniak and Johnson re-ranking parser.

CD-SCO corpus

Example sentence from the CD-SCO corpus

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	.	.	.	*)	_	_	_	_	_	_

CD-SCO corpus

Corpus statistics

	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

Evaluation

- The CoNLL 2010 ST introduced precision and recall at scope level as performance measures.
- The CONLL 2010 ST evaluation requirements were somewhat strict:
  - 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.
  - 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.

Evaluation

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

Evaluation

- 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).

Submissions

- Six teams (UiO1, UiO2, FBK, UWashington, UMichigan, UABCoRAL) submitted results for the closed track with a total of seven runs.
- Four teams (UiO2, UGroningen, UCM-1, UCM-2) submitted results for the open track with a total of five runs.

Results

		Cues	Scopes CM	Scopes NCM	Scope Tokens	Negated	Global	CNS
Closed track	**UiO1 r2**	91.31	70.39	70.39	82.37	**67.02**	**57.63**	**43.83**
	UiO1 r1	92.10	72.17	72.17	85.26	66.12	57.57	42.13
	UiO2	91.31	**72.39**	72.39	83.73	59.40	53.73	40.00
	FBK	**92.34**	70.39	70.39	81.98	60.20	50.93	35.74
	UWashington	90.00	71.81	**72.40**	83.51	54.25	48.09	34.04
	UMichigan	90.98	64.78	64.78	82.70	51.10	42.49	27.23
	UABCoRAL	85.77	63.46	64.76	76.23	48.33	39.04	26.81
Open track	**UiO2**	**91.31**	**72.39**	**72.39**	**82.20**	**61.79**	**54.82**	**41.28**
	UGroningen r2	86.82	53.26	53.26	75.17	60.65	39.56	27.23
	UCM-1	90.26	59.64	59.64	76.03	21.36	32.57	18.72
	UCM-2	71.88	48.98	49.24	62.65	29.03	17.46	11.91
	UGroningen r1	84.88	20.12	20.12	69.99	52.98	12.62	7.66

Best results

- Cue detection: FBK system (CRFs)
- Scope resolution: UWashington (CRFs) and UiO2 (CRFs)
- Negated events: UiO1(classification of factual events + SVM ranker)

Approaches

- 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 CRFs or 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.

Approaches

- The systems that participate in the closed track are machine learning based.
- The resources utilized by participants in the open track are diverse.
  - UiO2 reparsed the data with MaltParser in order to obtain dependency graphs.
  - The UGroningen system is based on tools that produce complex semantic representations: C&C tools for parsing and Boxer to produce semantic representations in the form of Discourse Representation Structures (DRSs).
  - UCM-1 and UCM-2 are rule-based systems that rely heavily on information from the syntax tree.

Analysis

Comparing results per track

- 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).

Comparing results per approach

- 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).

Analysis

Comparing results per substasks

- 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.
  - Error propagation
  - The set of negation cues is closed and comprises mostly single tokens
  - Scope sequences are longer

Outline

Task Description

- Resolving the focus of negation
  - only verbal, clausal and analytical negation
  - detecting negated statements is not part of this task
- Focus is either
  - the (negated) verb or
  - a semantic role of the verb
    the verb and other roles can be interpreted positive
- Why the full text of a role?
  - often times the focus can be narrowed down

PB-FOC corpus

- Annotation were done on top of PropBank
- Sentences marked with MNEG, total: 3,993
  - section 02-21 for training 24 for development and 23 for test 3,544 sentences
- Additional automatically obtained annotations:
  - Token number, POS tags, named entities, chunks, parse tree, dependency tree, semantic role labels, whether toke is negation

PB-FOC corpus

- Annotation were done on top of PropBank
- Sentences marked with MNEG, total: 3,993
  - section 02-21 for training 24 for development and 23 for test 3,544 sentences
- Additional automatically obtained annotations:
  - Token number, POS tags, named entities, chunks, parse tree, dependency tree, semantic role labels, whether toke is negation

Examples:

- *Even if that deal isn't {revived}, NBC hopes to find another.*
  Even if that deal is suppressed, NBC hopes to find another.

PB-FOC corpus

- Annotation were done on top of PropBank
- Sentences marked with MNEG, total: 3,993
  - section 02-21 for training 24 for development and 23 for test 3,544 sentences
- Additional automatically obtained annotations:
  - Token number, POS tags, named entities, chunks, parse tree, dependency tree, semantic role labels, whether toke is negation

Examples:

- *Even if that deal isn't {revived}, NBC hopes to find another.*
  Even if that deal is suppressed, NBC hopes to find another.
- *A decision isn't expected {until some time next year}.*
  A decision is expected at some time next year.

PB-FOC corpus

- Annotation were done on top of PropBank
- Sentences marked with MNEG, total: 3,993
  - section 02-21 for training 24 for development and 23 for test 3,544 sentences
- Additional automatically obtained annotations:
  - Token number, POS tags, named entities, chunks, parse tree, dependency tree, semantic role labels, whether toke is negation

Examples:

- *Even if that deal isn't {revived}, NBC hopes to find another.*
  Even if that deal is suppressed, NBC hopes to find another.
- *A decision isn't expected {until some time next year}.*
  A decision is expected at some time next year.
- *. . . it told the SEC it couldn't provide financial statements by the end of its first extension "{without unreasonable burden or expense}".*
  It could provide them by that time with a huge overhead.

PB-FOC corpus

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	*	*	-	*

PB-FOC corpus

		Train	Devel	Test
	1 role	2,210	515	672
	2 roles	89	15	38
	3 roles	3	0	2
	All	2,302	530	712
	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

Evaluation, participants and results

- Particpants are ranked by F-measure
  - perfect match
- One team participated, UConcordia; 2 runs, open track

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

Outline

Summing up

- We presented the description of the first *SEM Shared Task on Resolving the Scope and Focus of Negation.
- Two new datasets have been produced for this Shared Task: the CD-SCO corpus and the PB-FOC corpus.
- New evaluation software was also developed for this task.
- The number of submissions shows that there is interest in the topic within the computational linguistics community.

Future developments

- Unifying the annotation schemes of the two corpora.
- Annotating more data: financial and biomedical domain.
- Providing better evaluation measures for scope resolution.
- Going further: processing meanig at a deeper level.
  - Inferring implicit positive meaning.
  - Effect of negation processing in paraphrasing, summarization, textual entailment, etc.

Acknowledgements

- Thanks for your attention!
- And more thanks to:
  - Vivek Srikumar for pre-processing the PB-FOC corpus with the Illinois semantic role labeler.
  - Stephan Oepen for pre-processing the CD-SCO corpus.
  - The *SEM organisers and the ST participants.