Simple Negation Scope Resolution through Deep Parsing: A Semantic Solution to a Semantic Problem

Woodley Packard[♣], Emily M. Bender[♣], Jonathon Read[♠], Stephan Oepen[♡◇], and Rebecca Dridan[♡]

[♣] University of Washington, Department of Linguistics
[♠] Teesside University, School of Computing
[♡] University of Oslo, Department of Informatics
[◇] Potsdam University, Department of Linguistics

ebender@uw.edu, sweaglesw@sweaglesw.org, j.read@tees.ac.uk, {oe|rdridan}@ifi.uio.no

Abstract

In this work, we revisit Shared Task 1 from the 2012 *SEM Conference: the automated analysis of negation. Unlike the vast majority of participating systems in 2012, our approach works over explicit and formal representations of propositional semantics, i.e. derives the notion of negation *scope* assumed in this task from the structure of logical-form meaning representations. We relate the task-specific interpretation of (negation) scope to the concept of (quantifier and operator) scope in mainstream underspecified semantics. With reference to an explicit encoding of semantic predicate-argument structure, we can operationalize the annotation decisions made for the 2012 *SEM task, and demonstrate how a comparatively simple system for negation scope resolution can be built from an off-the-shelf deep parsing system. In a system combination setting, our approach improves over the best published results on this task to date.

1 Introduction

Recently, there has been increased community interest in the theoretical and practical analysis of what Morante and Sporleder (2012) call *modality and negation*, i.e. linguistic expressions that modulate the certainty or factuality of propositions. Automated analysis of such aspects of meaning is important for natural language processing tasks which need to consider the truth value of statements, such as for example text mining (Vincze et al., 2008) or sentiment analysis (Lapponi et al., 2012). Owing to its immediate utility in the curation of scholarly results, the analysis of negation and so-called hedges in bio-medical research literature has been the focus of several workshops, as well as the Shared Task at the 2011 Conference on Computational Language Learning (CoNLL).

Task 1 at the First Joint Conference on Lexical and Computational Semantics (*SEM 2012; Morante and Blanco, 2012) provided a fresh, principled annotation of negation and called for systems to analyze negation—detecting cues (affixes, words, or phrases that express negation), resolving their scopes (which parts of a sentence are actually negated), and identifying the negated event or property. The task organizers designed and documented an annotation scheme (Morante and Daelemans, 2012) and applied it to a little more than 100,000 tokens of running text by the novelist Sir Arthur Conan Doyle. While the task and annotations were framed from a semantic perspective, only one participating system actually employed explicit compositional semantics (Basile et al., 2012), with results ranking in the middle of the 12 participating systems. Conversely, the best-performing systems approached the task through machine learning or heuristic processing over *syntactic* and linguistically relatively *coarse-grained* representations; see § 2 below.

Example (1), where ⟨⟩ marks the cue and {} the in-scope elements, illustrates the annotations, including how negation inside a noun phrase can scope over discontinuous parts of the sentence.[1]

(1) {The German} was sent for but professed to {know} ⟨nothing⟩ {of the matter}.

In this work, we return to the 2012 *SEM task from a deliberately semantics-centered point of view, focusing on the hardest of the three sub-problems: scope resolution.[2] Where Morante and Daelemans (2012) characterize negation as an "extra-propositional aspect of meaning" (p. 1563),

[1] Our running example is a truncated variant of an item from the Shared Task training data. The remainder of the original sentence does not form part of the scope of this cue.

[2] Resolving negation scope is a more difficult sub-problem at least in part because (unlike cue and event identification) it is concerned with much larger, non-local and often discontinuous parts of each utterance. This intuition is confirmed by Read et al. (2012), who report results for each sub-problem using gold-standard inputs; in this setup, scope resolution showed by far the lowest performance levels.

we in fact see it as a core piece of compositionally constructed logical-form representations. Though the task-specific concept of scope of negation is not the same as the notion of quantifier and operator scope in mainstream underspecified semantics, we nonetheless find that reviewing the 2012 *SEM Shared Task annotations with reference to an explicit encoding of semantic predicate-argument structure suggests a simple and straightforward operationalization of their concept of negation scope. Our system implements these findings through a notion of functor-argument 'crawling', using as our starting point the underspecified logical-form meaning representations provided by a general-purpose deep parser.

Our contributions are three-fold: Theoretically, we correlate the structures at play in the Morante and Daelemans (2012) view on negation with formal semantic analyses; methodologically, we demonstrate how to approach the task in terms of underspecified, logical-form semantics; and practically, our combined system retroactively 'wins' the 2012 *SEM Shared Task. In the following sections, we review related work (§ 2), detail our own setup (§ 3), and present and discuss our experimental results (§ 4 and § 5, respectively).

2 Related Work

Read et al. (2012) describe the best-performing submission to Task 1 of the 2012 *SEM Conference. They investigated two approaches for scope resolution, both of which were based on syntactic constituents. Firstly, they created a set of 11 heuristics that describe the path from the preterminal of a cue to the constituent whose projection is predicted to match the scope. Secondly they trained an SVM ranker over candidate constituents, generated by following the path from a cue to the root of the tree and describing each candidate in terms of syntactic properties along the path and various surface features. Both approaches attempted to handle discontinuous instances by applying two heuristics to the predicted scope: (a) removing preceding conjuncts from the scope when the cue is in a conjoined phrase and (b) removing sentential adverbs from the scope. The ranking approach showed a modest advantage over the heuristics (with F_1 equal to 77.9 and 76.7, respectively, when resolving the scope of gold-standard cues in evaluation data). Read et al. (2012) noted however that the annotated scopes did not align with the Shared Task–provided constituents for 14% of the instances in the training data, giving an F_1 upper-bound of around 86.0 for systems that depend on those constituents.

Basile et al. (2012) present the only submission to Task 1 of the 2012 *SEM Conference which employed compositional semantics. Their scope resolution pipeline consisted primarily of the C&C parser and Boxer (Curran et al., 2007), which produce Discourse Representation Structures (DRSs). The DRSs represent negation explicitly, including representing other predications as being within the scope of negation. Basile et al. (2012) describe some amount of tailoring of the Boxer lexicon to include more of the Shared Task scope cues among those that produce the negation operator in the DRSs, but otherwise the system appears to directly take the notion of scope of negation from the DRS and project it out to the string, with one caveat: As with the logical-forms representations we use, the DRS logical forms do not include function words as predicates in the semantics. Since the Shared Task gold standard annotations included such arguably semantically vacuous (see Bender, 2013, p. 107) words in the scope, further heuristics are needed to repair the string-based annotations coming from the DRS-based system. Basile et al. resort to counting any words between in-scope tokens which are not themselves cues as in-scope. This simple heuristic raises their F_1 for full scopes from 20.1 to 53.3 on system-predicted cues.

3 System Description

The new system described here is what we call the MRS Crawler. This system operates over the normalized semantic representations provided by the LinGO English Resource Grammar (ERG; Flickinger, 2000).[3] The ERG maps surface strings to meaning representations in the format of Minimal Recursion Semantics (MRS; Copestake et al., 2005). MRS makes explicit predicate-argument relations, as well as partial information about scope (see below). We used the grammar together with one of its pre-packaged conditional Maximum Entropy models for parse ranking, trained on a combination of encyclopedia articles and tourism brochures. Thus, the deep parsing front-end system to our MRS Crawler has not been

[3] In our experiments, we use the 1212 release of the ERG, in combination with the ACE parser (`http://sweaglesw.org/linguistics/ace/`). The ERG and ACE are DELPH-IN resources; see `http://www.delph-in.net`.

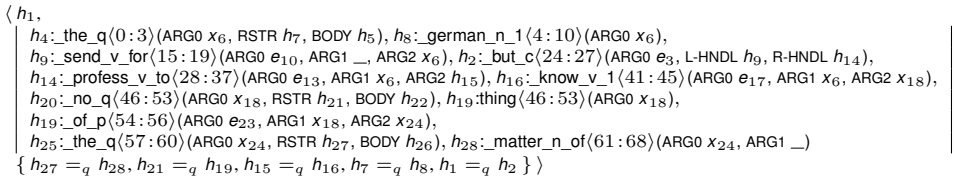

Figure 1: MRS analysis of our running example (1).

adapted to the task or its text type; it is applied in an 'off the shelf' setting. We combine our system with the outputs from the best-performing 2012 submission, the system of Read et al. (2012), firstly by relying on the latter for system negation cue detection,[4] and secondly as a fall-back in system combination as described in § 3.4 below.

Scopal information in MRS analyses delivered by the ERG fixes the scope of operators—such as negation, modals, scopal adverbs (including subordinating conjunctions like *while*), and clause-embedding verbs (e.g. *believe*)—based on their position in the constituent structure, while leaving the scope of quantifiers (e.g. *a* or *every*, but also other determiners) free. From these underspecified representations of possible scopal configurations, a scope resolution component can spell out the full range of fully-connected logical forms (Koller and Thater, 2005), but it turns out that such enumeration is not relevant here: the notion of scope encoded in the Shared Task annotations is not concerned with the relative scope of quantifiers and negation, such as the two possible readings of (2) represented informally below:[5]

(2) Everyone didn't leave.

a. $\forall(x)\neg\text{leave}(x) \sim$ Everyone stayed.

b. $\neg\forall(x)\text{leave}(x) \sim$ At least some stayed.

However, as shown below, the information about fixed scopal elements in an underspecified MRS is sufficient to model the Shared Task annotations.

3.1 MRS Crawling

Fig. 1 shows the ERG semantic analysis for our running example. The heart of the MRS is a multiset of elementary predications (EPs). Each elementary prediction includes a predicate symbol, a label (or 'handle', prefixed to predicates with a colon in Fig. 1), and one or more argument positions, whose values are semantic variables. Eventualities (e_i) in MRS denote states or activities, while instance variables (x_j) typically correspond to (referential or abstract) entities. All EPs have the argument position ARG0, called the *distinguished variable* (Oepen and Lønning, 2006), and no variable is the ARG0 of more than one non-quantifier EP.

The arguments of one EP are linked to the arguments of others either directly (sharing the same variable as their value), or indirectly (through so-called 'handle constraints', where $=_q$ in Fig. 1 denotes equality modulo quantifier insertion). Thus a well-formed MRS forms a connected graph. In addition, the grammar links the EPs to the elements of the surface string that give rise to them, via character offsets recorded in each EP (shown in angle brackets in Fig. 1). For the purposes of the present task, we take a negation cue as our entry point into the MRS graph (as our initial *active* EP), and then move through the graph according to the following simple operations to add EPs to the active set:

Argument Crawling Add to the scope all EPs whose distinguished variable or label is an argument of the active EP; for arguments of type h_k, treat any $=_q$ constraints as label equality.

Label Crawling Add all EPs whose label is identical to that of the active EP.

Functor Crawling Add all EPs that take the distinguished variable or label of the active EP as an argument (directly or via $=_q$ constraints).

Our MRS crawling algorithm is sketched in Fig. 2. To illustrate how the rules work, we will trace their operation in the analysis of example (1), i.e. traverse the EP graph in Fig. 1.

The negation cue is *nothing*, from character position 46 to 53. This leads us to _no_q as our entry point into the graph. Our algorithm states that for this type of cue (a quantifier) the first step is

[4] Read et al. (2012) predicted cues using a closed vocabulary assumption with a supervised classifier to disambiguate instances of cues.

[5] In other words, a possible semantic interpretation of the (string-based) Shared Task annotation guidelines and data is in terms of a quantifier-free approach to meaning representation, or in terms of one where quantifier scope need not be made explicit (as once suggested by, among others, Alshawi, 1992). From this interpretation, it follows that the notion of scope assumed in the Shared Task does not encompass interactions of negation operators and quantifiers.

1: Activate the cue EP
2: **if** the cue EP is a quantifier **then**
3: Activate EPs reached by functor crawling from the distinguished variable (ARG0) of the cue EP
4: **end if**
5: **repeat**
6: **for** each active EP X **do**
7: Activate EPs reached by argument crawling or label crawling unless they are co-modifiers of the negation cue.[a]
8: Activate EPs reached by functor crawling if they are modal verbs, or one of the following subordinating conjunctions reached by ARG1: *whether*, *when*, *because*, *to*, *with*, *although*, *unless*, *until*, or *as*.
9: **end for**
10: **until** a fixpoint is reached (no additional EPs were activated)
11: Deactivate zero-pronoun EPs (from imperative constructions)
12: Apply semantically empty word handling rules (iterate until a fixpoint is reached)
13: Apply punctuation heuristics

Figure 2: Algorithm for scope detection by MRS crawling

[a]Formally: If an EP shares its label with the negation cue, or is a quantifier whose restriction (RSTR) is $=_q$ equated with the label of the negation cue, it cannot be in-scope unless its ARG0 is an argument of the negation cue, or the ARG0 of the negation cue is one of its own arguments. See § 3.3 for elaboration.

functor crawling (see § 3.3 below), which brings _know_v_1 into the scope. We proceed with *argument crawling* and *label crawling*, which pick up _the_q⟨0:3⟩ and _german_n_1 as the ARG1. Further, as the ARG2 of _know_v_1, we reach thing and through recursive invocation we activate _of_p and, in yet another level of recursion, _the_q⟨57:60⟩ and _matter_n_of. At this point, crawling has no more links to follow. Thus, the MRS crawling operations 'paint' a subset of the MRS graph as in-scope for a given negation cue.

3.2 Semantically Empty Word Handling

Our crawling rules operate on semantic representations, but the annotations are with reference to the surface string. Accordingly, we need projection rules to map from the 'painted' MRS to the string. We can use the character offsets recorded in each EP to project the scope to the string. However, the string-based annotations also include words which the ERG treats as semantically vacuous. Thus in order to match the gold annotations, we define a set of heuristics for when to count vacuous words as in scope. In (1), there are no semantically empty words in-scope, so we illustrate these heuristics with another example:

(3) "I trust that {there is} ⟨nothing⟩ {of consequence which I have overlooked}?"

The MRS crawling operations discussed above paint the EPs corresponding to *is*, *thing*, *of*, *consequence*, *I*, and *overlooked* as in-scope (underlined in (3)). Conversely, the ERG treats the words *that*, *there*, *which*, and *have* as semantically empty. Of these, we need to add all except *that* to the scope. Our vacuous word handling rules use the syntactic structure provided by the ERG as scaffolding to help link the scope information gleaned from contentful words to vacuous words. Each node in the syntax tree is initially colored either in-scope or out-of-scope in agreement with the decision made by the crawler about the lexical head of the corresponding subtree. A semantically empty word is determined to be in-scope if there is an in-scope syntax tree node in the right position relative to it, as governed by a short list of templates organized by the type of the semantically empty word (particles, complementizers, non-referential pronouns, relative pronouns, and auxiliary verbs).

As an example, the rule for auxiliary verbs like *have* in our example (3) is that they are in scope when their verb phrase complement is in scope. Since *overlooked* is marked as in-scope by the crawler, the semantically empty *have* becomes in-scope as well. Sometimes the rules need to be iterated. For example, the main rule for relative pronouns is that they are in-scope when they fill a gap in an in-scope constituent; *which* fills a gap in the constituent *have overlooked*, but since *have* is the (syntactic) lexical head of that constituent, the verb phrase is not considered in-scope the first time the rules are tried.

Similar rules deal with *that* (complementizers are in-scope when the complement phrase is an argument of an in-scope verb, which is not the case here) and *there* (non-referential pronouns are in-scope when they are the subject of an in-scope VP, which is true here).

3.3 Re-Reading the Annotation Guidelines

Our MRS crawling algorithm was defined by looking at the annotated data rather than the annotation guidelines for the Shared Task (Morante et al., 2011). Nonetheless, our algorithm can be seen as a first pass formalization of the guidelines. In this section, we briefly sketch how our algorithm corresponds to different aspects of the guidelines.

For negated verbs, the guidelines state that "If the negated verb is the main verb in the sentence, the entire sentence is in scope." (Morante et al., 2011, 17). In terms of our operations defined over semantic representations, this is rendered as follows: all arguments of the negated verb are selected by *argument crawling*, all intersective modifiers by *label crawling*, and *functor crawling* (Fig. 2, line 8) captures modal auxiliaries and non-intersective modifiers. The guidelines treat predicative adjectives under a separate heading from verbs, but describe the same desired annotations (scope over the whole clause; ibid., p. 20). Since these structures are analogous in the semantic representations, the same operations that handle negated verbs also handle negated predicative adjectives correctly.

For negated subjects and objects, the guidelines state that the negation scopes over "all the clause" and "the clause headed by the verb" (Morante et al., 2011, 19), respectively. The examples given in the annotation guidelines suggest that these are in fact meant to refer to the same thing. The negation cue for a negated nominal argument will appear as a quantifier EP in the MRS, triggering line 3 of our algorithm. This *functor crawling* step will get to the verb's EP, and from there, the process is the same as the last two cases.

In contrast to subjects and objects, negation of a clausal argument is not treated as negation of the verb (ibid., p. 18). Since in this case, the negation cue will not be a quantifier in the MRS, there will be no *functor crawling* to the verb's EP.

For negated modifiers, the situation is somewhat more complex, and this is a case where our crawling algorithm, developed on the basis of the annotated data, does not align directly with the guidelines as given. The guidelines state that negated attributive adjectives have scope over the entire NP (including the determiner) (ibid., p. 20) and analogously negated adverbs have scope over the entire clause (ibid., p. 21). However, the annotations are not consistent, especially with respect to the treatment of negated adjectives: while the head noun and determiner (if present) are typically annotated as in scope, other co-modifiers, especially long, post-nominal modifiers (including relative clauses) are not necessarily included:

(4) "A dabbler in science, Mr. Holmes, a picker up of shells on the shores of {the} great ⟨un⟩{known ocean}.

(5) Our client looked down with a rueful face at {his} own ⟨un⟩{conventional appearance}.

(6) Here was {this} ⟨ir⟩{reproachable Englishman} ready to swear in any court of law that the accused was in the house all the time.

(7) {There is}, on the face of it, {something} ⟨un⟩{natural about this strange and sudden friendship between the young Spaniard and Scott Eccles}.

Furthermore, the guidelines treat relative clauses as subordinate clauses and thus negation inside a relative clause is treated as bound to that clause only, and includes neither the head noun of the relative clause nor any of its other dependents in its scope. However, from the perspective of MRS, a negated relative clause is indistinguishable from any other negated modifier of a noun. This treatment of relative clauses (as well as the inconsistencies in other forms of co-modification) is the reason for the exception noted at line 7 of Fig. 2. By disallowing the addition of EPs to the scope if they share the label of the negation cue but are not one of its arguments, we block the head noun's EP (and any EPs only reachable from it) in cases of relative clauses where the head verb inside the relative clause is negated. It also blocks co-modifiers like *great*, *own*, and the phrases headed by *ready* and *about* in (4)–(7). As illustrated in these examples, this is correct some but not all of the time. Having been unable to find a generalization capturing when comodifiers are annotated as in scope, we stuck with this approximation.

For negation within clausal modifiers of verbs, the annotation guidelines have further information, but again, our existing algorithm has the correct behavior: The guidelines state that a negation cue inside of the complement of a subordinating conjunction (e.g. *if*) has scope only over the subordinate clause (ibid., p. 18 and p. 26). The ERG treats all subordinating conjunctions as two-place predicates taking two scopal arguments. Thus, as with clausal complements of clause-embedding verbs, the embedding subordinating conjunction and any other arguments it might have are inaccessible, since functor crawling is restricted to a handful of specific configurations.

As is usually the case with exercises in formalization, our crawling algorithm generalizes beyond what is given explicitly in the annotation guidelines. For example, all arguments that are treated as semantically nominal (including PP arguments where the preposition is semantically null) are treated in the same way as subjects and objects; similarly, all arguments which are semantically clausal (including certain PP arguments) are handled the same way as clausal complements. This is possible because we take advantage of the high degree of normalization that the ERG accomplishes in mapping to the MRS representation.

There are also cases where we are more specific. The guidelines do not handle coordination in detail, except to state that in coordinated clauses negation is restricted to the clause it appears in (ibid., p. 17–18) and to include a few examples of coordination under the heading 'ellipsis'. In the case of VP coordination, our existing algorithm does not need any further elaboration to pick up the subject of the coordinated VP but not the non-negated conjunct, as shown in discussion of (1) in § 3.1 above. In the case of coordination of negated NPs, recall that to reach the main portion of the negated scope we must first apply functor crawling. The functor crawling procedure has a general mechanism to transparently continue crawling up through coordinated structures while blocking future crawling from traversing them again.[6]

On the other hand, there are some cases in the annotation guidelines which our algorithm does not yet handle. We have not yet provided any analysis of the special cases for *save* and *expect* discussed in Morante et al., 2011, pp. 22–23, and also do not have a means of picking out the overt verb in gapping constructions (p. 24).

Finally, we note that even carefully worked out annotation guidelines such as these are never followed perfectly consistently by the human annotators who apply them. Because our crawling algorithm so closely models the guidelines, this puts our system in an interesting position to provide feedback to the Shared Task organizers.

3.4 Fall-Back Configurations

The close match between our crawling algorithm and the annotation guidelines supported by the mapping to MRS provides for very high precision and recall when the analysis engine produces the desired MRS.[7] However, the analysis engine does not always provide the desired analysis, largely because of idiosyncrasies of the genre (e.g. vocatives appearing mid-sentence) that are either not handled by the grammar or not well modeled in the parse selection component. In addition, as noted above, there are a handful of negation cues we do not yet handle. Thus, we also tested fall-back configurations which use scope predictions based on MRS in some cases, and scope predictions from the system of Read et al. (2012) in others.

Our first fall-back configuration (Crawler$_N$ in Table 1) uses MRS-based predictions whenever there is a parse available and the cue is one that our system handles. Sometimes, the analysis picked by the ERG's statistical model is not the correct analysis for the given context. To combat such suboptimal parse selection performance, we investigated using the probability of the top ranked analysis (as determined by the parse selection model and conditioned on the sentence) as a confidence metric. Our second fall-back configuration (Crawler$_P$ in Table 1) uses MRS-based predictions when there is a parse available whose conditional probability is at least 0.5.[8]

4 Experiments

We evaluated the performance of our system using the Shared Task development and evaluation data (respectively CDD and CDE in Table 1). Since we do not attempt to perform cue detection, we report performance using gold cues and also using the system cues predicted by Read et al. (2012). We used the official Shared Task evaluation script to compute all scores.

4.1 Data Sets

The Shared Task data consists of chapters from the *Adventures of Sherlock Holmes* mystery novels and short stories. As such, the text is carefully edited turn-of-the-20th-century British English,[9]

[6] This allows *ate* to be reached in *We ate bread but no fish.*, while preventing *but* and *bread* from being reached, which they otherwise would via argument crawling from *ate*.

[7] And in fact, the task is somewhat noise-tolerant: some parse selection decisions are independent of each other, and a mistake in a part of the analysis far enough away from the negation cue does not harm performance.

[8] This threshold was determined empirically on the development data. We also experimented with other confidence metrics—the probability ratio of the top-ranked and second parse or the entropy over the probability distribution of the top 10 parses—but found no substantive differences.

[9] In contrast, the ERG was engineered for the analysis of contemporary American English, and an anecdotal analysis of parse failures and imperfect top-ranked parses suggests

		Gold Cues						System Cues					
		Scopes			Tokens			Scopes			Tokens		
Set	**Method**	**Prec**	**Rec**	**F_1**	**Prec**	**Rec**	**F_1**	**Prec**	**Rec**	**F_1**	**Prec**	**Rec**	**F_1**
CDD	Ranker	**100.0**	68.5	81.3	84.8	**86.8**	85.8	**91.7**	66.1	76.8	79.5	**84.9**	82.1
	Crawler	**100.0**	53.0	69.3	**89.3**	67.0	76.6	90.8	53.0	66.9	**84.7**	65.9	74.1
	Crawler$_N$	**100.0**	64.9	78.7	89.0	83.5	86.1	90.8	64.3	75.3	82.6	82.1	82.3
	Crawler$_P$	**100.0**	**70.2**	**82.5**	86.4	**86.8**	**86.6**	91.2	**67.9**	**77.8**	80.0	**84.9**	**82.4**
	Oracle	100.0	*76.8*	*86.9*	*91.5*	*89.1*	*90.3*						
CDE	Ranker	98.8	64.3	77.9	85.3	**90.7**	87.9	87.4	61.5	72.2	82.0	**88.8**	85.3
	Crawler	**100.0**	44.2	61.3	85.8	68.4	76.1	**87.8**	43.4	58.1	78.8	66.7	72.2
	Crawler$_N$	98.6	56.6	71.9	83.8	88.4	86.1	86.0	54.2	66.5	78.4	85.7	81.9
	Crawler$_P$	98.8	**65.5**	**78.7**	**86.1**	90.4	**88.2**	87.6	**62.7**	**73.1**	**82.6**	88.5	**85.4**
	Oracle	100.0	*70.3*	*82.6*	*89.5*	*93.1*	*91.3*						

Table 1: Scope resolution performance of various configurations over each subset of the Shared Task data. Ranker refers to the system of Read et al. (2012); Crawler refers to our current system in isolation, or falling back to the Ranker prediction either when the sentence is not covered by the parser (Crawler$_N$), or when the parse probability is predicted to be less than 0.5 (Crawler$_P$); finally, Oracle simulates best possible selection among the Ranker and Crawler predictions (and would be ill-defined on system cues).

annotated with token-level information about the cues and scopes in every negated sentence. The training set contains 848 negated sentences, the development set 144, and the evaluation set 235. As there can be multiple usages of negation in one sentence, this corresponds to 984, 173, and 264 instances, respectively.

Being rule-based, our system does not require any training data per se. However, the majority of our rule development and error analysis were performed against the designated training data. We used the designated development data for a single final round of error analysis and corrections. The system was declared frozen before running with the formal evaluation data. All numbers reported here reflect this frozen system.[10]

4.2 Results

Table 1 presents the results of our various configurations in terms of both (a) whole *scopes* (i.e. a true positive is only generated when the predicted scope matches the gold scope exactly) and (b) in-scope *tokens* (i.e. a true positive for every token the system correctly predicts to be in scope). The table also details the performance upper-bound for system combination, in which an oracle selects the system prediction which scores the greater token-wise F_1 for each gold cue.

The low recall levels for Crawler can be mostly attributed to imperfect parser coverage. Crawler$_N$, which falls back just for parse failure brings the recall back up, and results in F_1 levels closer to the system of Read et al. (2012), albeit still not quite advancing the state of the art (except over the development set). Our best results are from Crawler$_P$, which outperforms all other configurations on the development and evaluation sets.

The Oracle results are interesting because they show that there is much more to be gained in combining our semantics-based system with the Read et al. (2012) syntactically-focused system. Further analysis of these results to draw out the patterns of complementary errors and strengths is a promising avenue for future work.

4.3 Error Analysis

To shed more light on specific strengths and weaknesses of our approach, we performed a manual error analysis of scope predictions by Crawler, starting from gold cues so as to focus in-depth analysis on properties specific to scope resolution over MRSs. This analysis was performed on CDD, in order to not bar future work on this task. Of the 173 negation cue instances in CDD, Crawler by itself makes 94 scope predictions that exactly match the gold standard. In comparison, the system of Read et al. (2012) accomplishes 119 exact scope matches, of which 80 are shared with Crawler; in other words, there are 14 cue instances (or 8% of all cues) in which our approach can improve over the best-performing syntax-based submission to the original Shared Task.

that the archaic style in the 2012 *SEM Shared Task texts has a strong adverse effect on the parser.

[10]The code and data are available from `http://www.delph-in.net/crawler/`, for replicability (Fokkens et al., 2013).

We reviewed the 79 negation instances where Crawler made a wrong prediction in terms of exact scope match, categorizing the source of failure into five broad error types:

(1) Annotation Error In 11% of all instances, we consider the annotations erroneous or inconsistent. These judgments were made by two of the authors, who both were familiar with the annotation guidelines and conventions observable in the data. For example, Morante et al. (2011) unambiguously state that subordinating conjunctions shall not be in-scope (8), whereas relative pronouns should be (9), and a negated predicative argument to the copula must scope over the full clause (10):

(8) It was after nine this morning {when we} reached his house and {found} ⟨neither⟩ {you} ⟨nor⟩ {anyone else inside it}.

(9) "We can imagine that in the confusion of flight something precious, something which {he could} ⟨not⟩ {bear to part with}, had been left behind.

(10) He said little about the case, but from that little we gathered that he also was not ⟨dis⟩{satisfied} at the course of events.

(2) Parser Failure Close to 30% of Crawler failures reflect lacking coverage in the ERG parser, i.e. inputs for which the parser does not make available an analysis (within certain bounds on time and memory usage).[11] In this work, we have treated the ERG as an off-the-shelf system, but coverage could certainly be straightforwardly improved by adding analyses for phenomena particular to turn-of-the-20th-century British English.

(3) MRS Inadequacy Another 33% of our false scope predictions are Crawler-external, viz. owing to erroneous input MRSs due to imperfect disambiguation by the parser or other inadequacies in the parser output. Again, these judgments (assigning blame outside our own work) were double-checked by two authors, and we only counted MRS imperfections that actually involve the cue or in-scope elements. Here, we could anticipate improvements by training the parse ranker on in-domain data or otherwise adapting it to this task.

(4) Cue Selection In close to 9% of all cases, there is a valid MRS, but Crawler fails to pick out an initial EP that corresponds to the negation cue. This first type of genuine crawling failure often relates to cues expressed as affixation (11), as well as to rare usages of cue expressions that predominantly occur with different categories, e.g. *neither* as a generalized quantifier (12):

(11) Please arrange your thoughts and let me know, in their due sequence, exactly what those events are {which have sent you out} ⟨un⟩{brushed} and unkempt, with dress boots and waistcoat buttoned awry, in search of advice and assistance.

(12) You saw yourself {how} ⟨neither⟩ {of the inspectors dreamed of questioning his statement}, extraordinary as it was.

(5) Crawler Deficiency Finally, a little more than 16% of incorrect predictions we attribute to our crawling rules proper, where we see many instances of under-coverage of MRS elements (13, 14) and a few cases of extending the scope too wide (15). In the examples below, erroneous scope predictions by Crawler are indicated through underlining. Hardly any of the errors in this category, however, involve semantically vacuous tokens.

(13) He in turn had friends among the indoor servants who unite in {their} fear and ⟨dis⟩{like of their master}.

(14) He said little about the case, but from that little we gathered that {he also was} ⟨not⟩ {dissatisfied at the course of events}.

(15) I tell you, sir, {I could}n't move a finger, ⟨nor⟩ {get my breath}, till it whisked away and was gone.

		Scopes			Tokens		
	Method	**Prec**	**Rec**	**F_1**	**Prec**	**Rec**	**F_1**
CDE	Boxer	76.1	41.0	53.3	69.2	82.3	75.2
	Crawler	87.8	43.4	58.1	78.8	66.7	72.2
	Crawler$_P$	87.6	62.7	73.1	82.6	88.5	85.4

Table 2: Comparison to Basile et al. (2012).

5 Discussion and Comparison

The example in (1) nicely illustrates the strengths of the MRS Crawler and of the abstraction provided by the deep linguistic analysis made possible by the ERG. The negated verb in that sentence is *know*, and its first semantic argument is *The German*. This semantic dependency is directly and explicitly represented in the MRS, but the phrase expressing the dependent is not adjacent to the head in the string. Furthermore, even a system using syntactic structure to model scope would be faced with a more complicated task than our crawling rules: At the level of syntax the dependency is mediated by both verb phrase coordination and the control verb *profess*, as well as by the semantically empty infinitival marker *to*.

[11] Overall parsing coverage on this data is about 86%, but of course all parser failures on sentences containing negation surface in our error analysis of Crawler in isolation.

The system we propose is very similar in spirit to that of Basile et al. (2012). Both systems map from logical forms with explicit representations of scope of negation out to string-based annotations in the format provided by the Shared Task gold standard. The main points of difference are in the robustness of the system and in the degree of tailoring of both the rules for determining scope on the logical form level and the rules for handling semantically vacuous elements. The system description in Basile et al. (2012) suggests relatively little tailoring at either level: aside from adjustments to the Boxer lexicon to make more negation cues take the form of the negation operator in the DRS, the notion of scope is directly that given in the DRS. Similarly, their heuristic for picking up semantically vacuous words is string-based and straightforward. Our system, on the other hand, models the annotation guidelines more closely in the definition of the MRS crawling rules, and has more elaborated rules for handling semantically empty words. The Crawler alone is less robust than the Boxer-based system, returning no output for 29% of the cues in CDE. These factors all point to higher precision and lower recall for the Crawler compared to the Boxer-based system. At the token level, that is what we see. Since full-scope recall depends on token-level precision, the Crawler does better across the board at the full-scope level. A comparison of the results is shown in Table 2.

A final key difference between our results and those of Basile et al. (2012) is the cascading with a fall-back system. Presumably a similar system combination strategy could be pursued with the Boxer-based system in place of the Crawler.

6 Conclusion and Outlook

Our motivation in this work was to take the design of the 2012 *SEM Shared Task on negation analysis at face value—as an overtly *semantic* problem that takes a central role in our long-term pursuit of *language understanding*. Through both theoretical and practical reflection on the nature of representations at play in this task, we believe we have demonstrated that explicit semantic structure will be a key driver of further progress in the analysis of negation. We were able to closely align two independently developed semantic analyses—the negation-specific annotations of Morante et al. (2011), on the one hand, and the broad-coverage, MRS meaning representations of the ERG, on the other hand. In our view, the conceptual correlation between these two semantic views on negation analysis reinforces their credibility.

Unlike the rather complex top-performing systems from the original 2012 competition, our MRS Crawler is defined by a small set of general rules that operate over general-purpose, explicit meaning representations. Thus, our approach scores high on transparency, adaptability, and replicability. In isolation, the Crawler provides premium precision but comparatively low recall. Its limitations, we conjecture, reflect primarily on ERG parsing challenges and inconsistencies in the target data. In a sense, our approach pushes a larger proportion of the task into the parser, meaning (a) there should be good opportunities for parser adaptation to this somewhat idiosyncratic text type; (b) our results can serve to offer feedback on ERG semantic analyses and parse ranking; and (c) there is a much smaller proportion of very task-specific engineering. When embedded in a confidence-thresholded cascading architecture, our system advances the state of the art on this task, and oracle combination scores suggest there is much remaining room to better exploit the complementarity of approaches in our study. In future work, we will seek to better understand the division of labor between the systems involved through contrastive error analysis and possibly another oracle experiment, constructing gold-standard MRSs for part of the data. It would also be interesting to try a task-specific adaptation of the ERG parse ranking model, for example retraining on the pre-existing treebanks but giving preference to analyses that lead to correct Crawler results downstream.

Acknowledgments

We are grateful to Dan Flickinger, the main developer of the ERG, for many enlightening discussions and continuous assistance in working with the analyses available from the grammar. This work grew out of a discussion with colleagues of the Language Technology Group at the University of Oslo, notably Elisabeth Lien and Jan Tore Lønning, to whom we are indebted for stimulating cooperation. Furthermore, we have benefited from comments by participants of the 2013 DELPH-IN Summit, in particular Joshua Crowgey, Guy Emerson, Glenn Slayden, Sanghoun Song, and Rui Wang.

References

Alshawi, H. (Ed.). 1992. *The Core Language Engine*. Cambridge, MA, USA: MIT Press.

Basile, V., Bos, J., Evang, K., and Venhuizen, N. 2012. UGroningen. Negation detection with Discourse Representation Structures. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics* (p. 301 – 309). Montréal, Canada.

Bender, E. M. 2013. *Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax*. San Rafael, CA, USA: Morgan & Claypool Publishers.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. 2005. Minimal Recursion Semantics. An introduction. *Research on Language and Computation*, *3*(4), 281 – 332.

Curran, J., Clark, S., and Bos, J. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Meeting of the Association for Computational Linguistics Demo and Poster Sessions* (p. 33 – 36). Prague, Czech Republic.

Flickinger, D. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, *6 (1)*, 15 – 28.

Fokkens, A., van Erp, M., Postma, M., Pedersen, T., Vossen, P., and Freire, N. 2013. Offspring from reproduction problems. What replication failure teaches us. In *Proceedings of the 51th Meeting of the Association for Computational Linguistics* (p. 1691 – 1701). Sofia, Bulgaria.

Koller, A., and Thater, S. 2005. Efficient solving and exploration of scope ambiguities. In *Proceedings of the 43rd Meeting of the Association for Computational Linguistics: Interactive Poster and Demonstration Sessions* (p. 9 – 12). Ann Arbor, MI, USA.

Lapponi, E., Read, J., and Øvrelid, L. 2012. Representing and resolving negation for sentiment analysis. In *Proceedings of the 2012 ICDM workshop on sentiment elicitation from natural text for information retrieval and extraction.* Brussels, Belgium.

Morante, R., and Blanco, E. 2012. *SEM 2012 Shared Task. Resolving the scope and focus of negation. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics* (p. 265 – 274). Montréal, Canada.

Morante, R., and Daelemans, W. 2012. ConanDoyle-neg. Annotation of negation in Conan Doyle stories. In *Proceedings of the 8th International Conference on Language Resources and Evaluation.* Istanbul, Turkey.

Morante, R., Schrauwen, S., and Daelemans, W. 2011. *Annotation of negation cues and their scope guidelines v1.0* (Tech. Rep. # CTRS-003). Antwerp, Belgium: Computational Linguistics & Psycholinguistics Research Center, Universiteit Antwerpen.

Morante, R., and Sporleder, C. 2012. Modality and negation. An introduction to the special issue. *Computational Linguistics*, *38*(2), 223 – 260.

Oepen, S., and Lønning, J. T. 2006. Discriminant-based MRS banking. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (p. 1250 – 1255). Genoa, Italy.

Read, J., Velldal, E., Øvrelid, L., and Oepen, S. 2012. UiO1. Constituent-based discriminative ranking for negation resolution. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics* (p. 310 – 318). Montréal, Canada.

Vincze, V., Szarvas, G., Farkas, R., Móra, G., and Csirik, J. 2008. The BioScope corpus. Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, *9(Suppl 11)*.