This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402.

Health in my Language

HimL

D2.1: Initial report on employing semantic role labelling and fidelity checking in machine translation

Author(s): Ondřej Bojar, Federico Fancellu, Jindřich Helcl, Phil Williams

Dissemination Level: Public

Date: February 1st 2016

Version 1.0

Grant agreement no.	644402
Project acronym	HimL
Project full title	Health in my Language
Funding Scheme	Innovation Action
Coordinator	Barry Haddow (UEDIN)
Start date, duration	1 February 2015, 36 months
Distribution	Public
Contractual date of delivery	February 1st 2016
Actual date of delivery	February 1st 2016
Deliverable number	D2.1
Deliverable title	Initial report on employing semantic role labelling and fidelity checking in machine translation
Type	Report
Status and version	1.0
Number of pages	37
Contributing partners	CUNI, UEDIN
WP leader	CUNI
Task leader	CUNI
Authors	Ondřej Bojar, Federico Fancellu, Jindřich Helcl, Phil Williams
EC project officer	Martina Eydner
The Partners in HimL are:	The University of Edinburgh (UEDIN), United Kingdom
	Univerzita Karlova V Praze (CUNI), Czech Republic
	Ludwig-Maximilians-Universitaet Muenchen (LMU-MUENCHEN), Germany
	Lingea SRO (LINGEA), Czech Republic
	NHS 24 (Scotland) (NHS24), United Kingdom
	Cochrane (COCHRANE), United Kingdom

For copies or reports, updates on project activities and other HimL-related information, contact:

Barry Haddow bhaddow@staffmail.ed.ac.uk
University of Edinburgh Phone: +44 (0) 131 651 3173

Contents

Executive Summary

The goal of WP2 Semantically Motivated MT is to improve the accuracy of statistical machine translation. We address the two frequent error types in the first two tasks: T2.1 aimed at preserving semantic roles of verb complements and modifiers, and T2.2 aimed at correct handling of negation.

Two additional tasks complement these efforts in a more direct way: T2.3 aims at avoiding critical translation errors of phrase-based (and hierarchical) model by removing translation units that are likely to introduce erros, and T2.4 will try to make MT benefit from existing large-scale high-quality dictionaries developed by Lingea.

The status of the tasks at the end of M12 is summarized in the following table:

Task	Months	Status
2.1: Modelling Semantic Role Labelling in MT	1–24	AS PLANNED
2.2: Enforcing Negation through Shallow Semantics	13–26	STARTED EARLIER
2.3: Improving Core Fidelity of Shallow Models	1–24	AS PLANNED
2.4: Employing Dictionaries	13–36	NOT STARTED YET

Details on the progress and experiments in the individual tasks are provided in the sections below.

1 Task 2.1: Modelling Semantic Role Labelling in Machine Translation

Semantic roles specify the relationships between the verbs of a sentence and their arguments. For instance, in the sentence 'John likes Mary,' the arguments 'John' and 'Mary' have different roles: John is the person who loves and Mary is the person who is loved. These distinct roles are often given generalized labels like 'agent' (for the person or thing performing the action of the verb) and 'patient' (for the person or thing to whom the action of the verb is directed). For a translation to faithfully convey the meaning of a source text, it is crucial that the semantic roles of the source text are correctly mapped to the target language. For instance, 'John loves Mary' can be correctly translated to the German 'John liebt Mary,' but if during translation the phrases 'John' and 'Mary' are reordered, then their semantic roles will be flipped. Since most arguments are translated independently, the only defence against unwanted changes of meaning are the reordering model and the *n*-gram language model, both of which are somewhat limited with respect to semantics. Unsurprisingly, changes in meaning are all too common in the output of real translation systems.

Most statistical machine translation models have no explicit concept of semantics, although there is now a small but growing body of literature on modelling semantic roles in machine translation. To date, the majority of this work has focussed on Chinese-English translation (among other reasons, this partly reflects the availability of suitably annotated corpora and language processing tools for both of those languages). In this project, our primary approach to modelling semantic roles builds on the work of Li *et al.* (2013) who introduce an predicate-argument structure model. Their model is language-independent and requires only source-side semantic annotation. As a result, it can be readily applied to the HimL language pairs. In the following section, we will describe Li *et al.* (2013)'s approach, our (partial) reimplementation, and our preliminary results.

A prerequisite for the implementation of Li *et al.* (2013)'s model is the extension of the Moses toolkit to include support for the representation and use of semantic roles in the translation pipeline. A happy side effect of this engineering work is that it opens up the toolkit for the implementation of alternative approaches to modelling semantic roles. Since it is not clear that one approach will be optimal for all HimL language pairs, it is important that there is scope for exploration and comparison of alternatives. We will briefly discuss some of these alternative approaches, including that of Bazrafshan and Gildea (2013), for which we also have preliminary results.

1.1 PropBank-style Semantic Role Labelling

The development of automatic semantic role labellers requires the provision of manually annotated training data. This data is expensive to produce and consequently only a few such data sources exist. These include the annotation efforts of PropBank (for English and Chinese), FrameNet (for English) and Abstract Meaning Representation (for English), each of which uses a different role set (with different linguistic underpinnings). Czech is also very well supported with linguistic resources, cf. Prague treebanks (Hajič *et al.*, 2012; Hajič *et al.*, 2006) and valency dictionaries (Lopatková *et al.*, 2006; Urešová *et al.*, 2016). We expect some resources to be available also for German but less so for Polish and Romanian. Although Li *et al.* (2013)'s model is somewhat neutral with regard to the role set, in experiments they use PropBank and so we follow them in doing so here.

PropBank, like most modern role sets, is defined on a verb-by-verb basis. Each distinct sense of a verb has an associated frameset, which specifies the number of possible arguments of a verb and their individual roles. As an example, Figure 1 gives

Frameset **open.01** "cause to open"
Arg0: agent
Arg1: thing opened
Arg2: instrument
Ex1: [$_{Arg0}$ John] *opened* [$_{Arg1}$ the door]
Ex2: [$_{Arg1}$ The door] *opened*
Ex3: [$_{Arg0}$ John] *opened* [$_{Arg1}$ the door] [$_{Arg2}$ with his foot]

Figure 1: The PropBank frameset for the verb 'open' with the sense 'cause to open'.

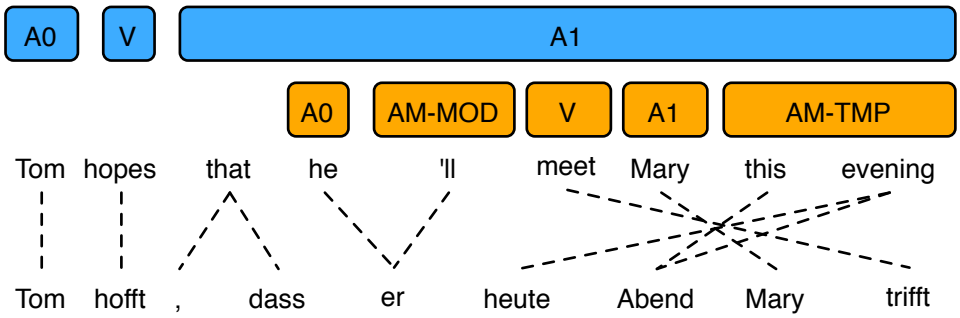

Figure 2: A word-aligned German-English sentence pair from the training data. The two semantic frames of the English source sentence are indicated by the blue and orange boxes above. Each box spans the words of a frame element. For instance, the first frame has a verb and two arguments, labelled A0 and A1. The A0 argument is 'Tom' and the A1 argument is 'that he'll meet Mary this evening'.

the frameset for the verb 'open' with the sense intended in a sentence like 'she opened the door' (as opposed to, for instance, the sense used in 'she opened a restaurant'). The frameset specifies that there are three potential arguments to the verb, labelled Arg0, Arg1, and Arg2. For most verbs, Arg0 is the agent (or a loosely agent-like role) and Arg1 is the patient (or a loosely patient-like role). Other numbered arguments (Arg2, Arg3, and so on) have verb-specific roles that are harder to generalize. These numbered arguments are referred to as core arguments. In addition, there are roles for non-core modifier arguments, such as AM-MOD (modal), AM-TMP (temporal), and AM-NEG (negation).

Figure 2 gives an example of a word-aligned sentence pair in which the English source sentence is annotated with semantic role labels. The verb 'hopes' has two arguments, A0 (which is the agent 'Tom') and A1, the thing that is hoped for. The verb 'meet' has two core arguments, A0 and A1, and two non-core arguments, AM-MOD and AM-TMP. We will use this annotated sentence pair as a running example in the description of this task.

1.1.1 Support for PropBank-style Semantic Role Labelling in Moses

We have extended the Moses toolkit to include some support for PropBank-style annotation. Li *et al.* (2013) use the Berkeley parser and an in-house semantic role labeller in their experiments. Currently, we assume the use of the SENNA toolkit (Collobert *et al.*, 2011) for both parsing and semantic role labelling. Specifically, Moses now has:

- A wrapper for SENNA that produces 1-best syntactic parse trees in Moses XML format, annotated with semantic role information. A fragment of the resulting XML for our English example sentence is given in Figure 3.
- Support in the rule extraction and scoring pipeline for accessing and making use of semantic role information.

When evaluated on standard tasks, the SENNA toolkit has been shown to produce results that are close to state of the art (Collobert *et al.*, 2011). However, for the current application it has a few disadvantages:

- SENNA can only produce 1-best annotations. In some models (as we will describe later), it may be useful to use the *k*-best trees (and corresponding semantic role labels).
- If the SENNA parse trees are used in syntax-based models, then translation quality is slightly degraded compared to the Berkeley parser or Brown parser.

```
...
<tree label="VP">
  <tree label="VBZ" semantic-frame-0="V"> hopes </tree>
  <tree label="SBAR" semantic-frame-0="A1">
    <tree label="IN"> that </tree>
    <tree label="S">
      <tree label="NP" semantic-frame-1="A0">
        <tree label="PRP"> he </tree>
      </tree>
      ...
```

Figure 3: A fragment of the source-side parse tree with semantic role annotation (Moses XML format).

- The semantic role labels and parse trees are produced independently. Since an argument must be a constituent, the two annotations cannot easily be aligned if they have different argument boundaries. Currently, such frames are simply discarded.

As an alternative to SENNA, we could also move from PropBank-style semantic roles labels to "functors" in the tectogrammatical representation of sentences, see the Prague Czech-English Dependency Treebank (Hajič *et al.*, 2012) for a brief description and references to the underlying theory, and use the automatic functor assignment tool by Ondřej Dušek as developed for the project FAUST[1]. The main change since the Deliverable is that the tool now covers not only Czech but also English and that it relies on Vowpal Wabbit learning toolkit. The current state of the art in semantic role labelling involves standard machine learning techniques. In brief, features are extracted from a parse tree and given as input to a classifier which determines for each tree node whether or not it is part of a verb or argument and, if so, what its label is.

1.2 Li et al's (2013) Semantic Role Mapping Model

Li *et al.* (2013) extend the hierarchical phrase-based model Chiang (2007) to include a model for mapping semantic roles from the source sentence to target sentence. Their model allows for the reordering and deletion of arguments. The Moses toolkit has included support for hierarchical phrase-based training and decoding for a number of years. Li *et al.* (2013)'s model involves the following changes to the standard hierarchical phrase based model:

1. Rule extraction uses hard syntactic constraints. For each sentence pair, a source-side parse tree is used to restrict the spans from which initial phrases are extracted. To avoid being over-restrictive, both the parse tree and extraction process are relaxed.
2. Semantic roles are mapped from the source side of the training data to the target side, via word alignments.
3. A semantic role mapping model is learned from the training data. This is referred to as the Predicate-Argument Structure (PAS) reordering model.
4. Decoding uses an input parse tree to close, i.e. prohibit, chart cells. In other words, rules can be applied over some spans and not others, as determined by the parse tree.
5. The PAS model is used as a feature function during decoding.

So far, we have implemented items 1, 2, and 4. We will give some details shortly, along with some preliminary results and analysis on argument reordering patterns in several language pairs.

1.2.1 Syntactic Constraints: Implementation

Li *et al.* (2013) provide an in-depth description of their syntactic constraints. We have reproduced their method without change. In brief:

1. The head words of nodes are marked (we implement Michael Collins' head finding rules (Collins, 1999)) and the trees are flattened according to head values. If a node m and a child node n have the same head word, then the child node n is eliminated from m's list of children and replaced by n's children.

[1] ftp://mi.eng.cam.ac.uk/pub/faust-pub/Deliverables/FAUSTD5.5.pdf, page 11

	en-de			en-ro		en-zh
	Cochrane	NHS24	Khresmoi	Cochrane	NHS24	NewsCommentary
Phrase-based	35.5	28.0	18.4	34.0	27.1	11.8
Hierarchical phrase-based	34.7	27.4	17.9	32.6	27.3	11.9
Tree-to-string	33.1	24.6	19.2	32.1	21.0	12.3
Li *et al.* (2013) constraints	33.3	26.1	18.2	31.1	26.1	12.1

Table 1: Translation quality (averaged BLEU) for baselines and systems using hard syntactic constraints.

2. Rule extraction is allowed for any span in which there is i) a single constituent, or ii) two or more sibling constituents (after tree flattening).
3. The limit on initial phrase length is removed (allowing rules that span the entire sentence).

During decoding the same span constraints used in item 2 are also used to determine which chart cells are closed.

1.2.2 Syntactic Constraints: Preliminary Results and Analysis

In our preliminary experiments, we test the effect of using syntactic constraints. Li *et al.* (2013) report Chinese-to-English results for a comparable experiment using three newswire test sets. For two test sets, they observe a significant improvement in BLEU and for the third, the difference is negligible. On average, they observe an improvement of 0.6 BLEU.

In these experiment, we have used two of the four HimL language pairs, English-German and English-Romanian. Since the original work that we seek to reproduce (and extend) was performed on the English-Chinese language pair, we also include that language pair as a point of comparison (although in the reverse direction, since we do not have a Chinese semantic role labeller).

As baselines we built phrase-based, hierarchical phrase-based models, and tree-to-string models (for further details of the last model type, see Williams *et al.* (2015)). We sampled 2M sentence pairs of training data for each language pair (as in the HimL systems, we used data from the OPUS repository for the Chinese-English system). Apart from minor details (and the reduced training data size), the phrase-based baselines are configured the same as the Y1 systems.

We compare the baselines against a hierarchical phrase-based model with hard syntactic constraints (as described above). Table 1 gives the BLEU scores, which are averaged over three tuning runs. The results are somewhat mixed and will require further analysis. For all of the HimL test sets (Cochrane and NHS24), using parse trees (either Li et al's constraints or tree-to-string) leads to a drop in translation quality compared to the phrase-based or hierarchical phrase-based model. In the other two cases (Khresmoi for English-German and NewsCommentary for English-Chinese), there is a small improvement from using hard constraints.

One particularly surprising result is the drop of 6 BLEU points on the NHS24 test set for the English-Romanian tree-to-string system. Examining the translations reveals that this is due to the decoder frequently choosing to use a rule that appends a sequence of underscores to a sentence, resulting in translation like the following:

```
healthy bones - Falls prevention &bar; NHS inform
s??n??tatea oaselor - USA prevenirea &bar; NHS cu _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
```

This rule reflects a pattern that occurs occasionally in the training data, but that is undesirable when translating the test set. Simply stripping the underscores boosts the BLEU score by 5.4 points. Further inspection of the 1-best output indicates that there are similar rules being used out of context and that there is likely to be scope for improving tree-to-string translation quality through the use of additional features that regulate the introduction of unaligned target words.

It is currently unclear how much of the BLEU scores variability for the different model types is to do with linguistic qualities of the languages, how much is to do with parse quality, and how much it is to do with the domain or quirks of the training data. We will explore these questions in subsequent work.

1.2.3 Semantic Role Mapping: Implementation

Li *et al.* (2013) do not provide details of their semantic role mapping method, other than to say that roles are projected across word alignments. This leaves some open questions. For instance, in Figure 2 what is the projected role label for 'er'? Is it A0, AM-MOD, or something else?

Our current implementation takes as input a word-aligned parallel corpus with source-side role annotation. For each sentence pair, it outputs a label sequence for each frame. For our example sentence pair, it outputs the following two sequences:

	en-de	en-ro	en-zh
Normalized	0.86	0.92	0.88
Adjusted	0.79	0.89	0.84

Table 2: Kendall-Tau distance for argument permutations.

```
hopes ||| A0 V A1 A1 A1 A1 A1 A1 A1 ||| A0 V A1 A1 A1 A1 A1 A1 A1
meet ||| - - - A0 AM-MOD V A1 AM-TMP AM-TMP ||| - - - - - A0 AM-TMP AM-TMP A1 V
```

The first field is the verb, the second is the source label sequence, and the third is the target label sequence.

If a target word has multiple projected role labels then the count of each label is recorded and the most frequent label is chosen. If there is a tie then the following disambiguation method is used:

- In a first pass over the data, we gather the projected role label counts for each target word type (in the example, we count how many times A0, AM-MOD, and any other label, is projected onto the word 'er').
- During the second pass, we pick the candidate label that has been observed most frequently for the given target word.

The resulting sequence is then reduced by eliminating non-labels and merging contiguous repeated labels. For the example, this yields:

```
hopes ||| A0 V A1 ||| A0 V A1
meet ||| A0 AM-MOD V A1 AM-TMP ||| A0 AM-TMP A1 V
```

If the resulting target label sequence contains repeated occurrences of a label (e.g. A0 V A0) then the frame is deemed inconsistent and discarded.

1.2.4 Semantic Role Mapping: Preliminary Results and Analysis

We have not yet implemented Li *et al.* (2013)'s PAS model, but to give a sense of how much predicate-argument reordering there is for different language pairs (and therefore how much scope the model has for influencing translation), we ran our semantic role mapper over the 2M word-aligned sentence pairs for the three language pairs used in the previous experiments. We ignored deleted arguments and focussed on reordering. For our example sentence pair, this measures the degree of reordering in the following pairs of role label sequences:

```
hopes ||| A0 V A1 ||| A0 V A1
meet ||| A0 V A1 AM-TMP ||| A0 AM-TMP A1 V
```

To measure reordering we calculated the Kendall-Tau distance for each semantic frame and averaged it across the training set. The Kendall-Tau distance has been used previously in machine translation for measuring reordering at the word level (Birch and Osborne, 2011). We calculate both the normalized Kendall-Tau distance and, following Birch and Osborne (2011), the adjusted Kendall-Tau distance (the square root of the standard metric). We subtract both from 1, so that a value of 0 indicates maximum disagreement and 1 indicates none.

Results are give in Table 2. These results suggests that there is a higher degree of predicate-argument reordering in English-German than English-Chinese, which bodes well for the application of the PAS model to that language pair. The higher value for English-Romanian suggests that there will be less scope for improvement in that language pair.

1.3 Alternative Models

Adding general support for semantic role labels in Moses has reduced the engineering effort required to experiment with alternative models. So far, we have re-implemented one approach, which we will briefly describe here, along with some preliminary results.

1.3.1 Bazrafshan and Gildea (2013)

Bazrafshan and Gildea (2013) extend the GHKM rule extraction method (Galley *et al.*, 2004) to take semantic role structure into account. Their approach involves running a semantic role labeller over the target-side of the training data and then performing

	System	test 1	test 2
English → German	Hiero	17.6	20.0
	T2S	18.0	20.3
	+ SRL	18.1 +0.1	20.4 +0.1
English → Russian	T2S	29.9	25.8
	+ SRL	30.0 +0.1	26.0 +0.2
Russian → English	S2T	28.1	22.8
	+ SRL	28.3 +0.2	23.1 +0.3
English → Czech	T2S	21.8	
	+ SRL	21.8 –	

Table 3: Preliminary results for our reimplementation of Bazrafshan and Gildea (2013)'s model.

a modified two-pass rule extraction step: in the first pass, standard GHKM rules are extracted; in the second, 'semantically-complete' rules are extracted. These must include either all arguments of a predicate or none. An indicator feature is used to distinguish the two types of rule. The rules are then used in standard string-to-tree decoding.

We have reimplemented their method and run some preliminary experiments. These experiments were run before the HimL systems were created and so are based on the systems built for the WMT15 translation task (Williams *et al.*, 2015). We use SENNA for semantic role labeller (as opposed to Bazrafshan and Gildea (2013)'s in-house labeller)

Results are given in Table 3. There is a consistent, albeit very small, increase in translation quality.

In a preliminary analysis, we manually inspected English-to-German translations, using sentence-level BLEU scores as a guide to identifying where the model was having an effect. This analysis suggested that the main effects were that there was more verb movement and less verb deletion. We intend to re-run these experiments on the HimL systems and to conduct a more detailed analysis.

1.4 Conclusion

We have described our progress so far on modelling semantic role labelling in machine translation. This includes the partial reimplementation of Li *et al.* (2013)'s model and the reimplementation of Bazrafshan and Gildea (2013)'s model. The results so far have been mixed and require further analysis. Our preliminary analysis of predicate-argument reordering suggests that there is particular scope for improvement through using Li *et al.* (2013)'s PAS model for the English-German language pair. We have found that using hard syntactic constraints does not always improve translation quality, although this may change as we add features to improve the models.

We intend to continue using an experiment-led approach to determine what works best for the HimL systems. One avenue for improving the use of hard syntactic constraints is using *k*-best or forest-based methods instead of relying on 1-best trees (Mi *et al.*, 2008). Since SENNA does not produce *k*-best or forest outputs, there are two options: to use an approach based on transformation of the 1-best output (e.g. Zhang *et al.* (2011)) or to develop an in-house labeller. We will explore the feasibility of both options.

2 Task 2.2: Enforcing Negation Through Shallow Semantics

The goal of Task 2.2 is to ensure that polarity (negation) is well preserved in translation, with a focus on the medical domain. The task is planned to start in year 2 of the project but UEDIN within HimL have already done significant work in this direction. A full description of the negation related project is included in Appendix A.

The goal of handling negation is also partly covered by Task 2.3 (Section 3) and Task 3.3 (Corrective approaches to morphology, see Deliverable 3.1).

The work on Task 2.2 we have done during the first year of HimL can be summarized as follows:

1. Getting acquainted with previous work on the topic.
2. Understanding where the gaps in previous work are: Through an analysis on what has been done with regard to negation in SMT, we can conclude that:
    - Negation is only considered in a few works and it is often treated as a side problem. Collins *et al.* (2005) and Li *et al.* (2009) consider it among other linguistic phenomena in the bigger picture of clause restructuring and re-ordering; Baker *et al.* (2012) take it into consideration only when it is associated with modality; finally Popovic and

Arcan (2015) and Bojar *et al.* (2013) take negation into account only amongst other types of errors the SMT system produces.

- Even when negation is the focus of a study, the scope is narrow and only one hypothesis or language pair is considered.
- None of the previous studies has conducted an extensive error analysis on the phenomenon so it is not clear what exactly are the kind of errors produced by an SMT system and what those depend on. This is connected to the previous item, where an hypothesis is formulated without any thorough inspection of the data.

3. Understanding negation, with particular emphasis of its translation: Analysing the errors involved in translating negation leads us to the question of what is meant by negation and what are we assessing the translation of. Previous work that have specifically addressed the problem of translating negation (Wetzel and Bond (2012) and Bojar *et al.* (2013) in particular) have cast the problem of translating negation as a binary outcome of whether a single element, the negation marker, has been translated or dropped during translation. Reducing it to a presence vs. absence problem is however an oversimplification. The negative marker interacts in fact with certain elements in the sentence whilst some others are outside its scope; this implies that if during translation, certain elements that are not under the scope of negation are brought inside it or viceversa, the overall rendering of the negation phenomenon is not correct. For this reason, the present work starts from the assumption that negation is a structural phenomenon, composed of a finite set of elements that interact with each other. Following Blanco and Moldovan (2011), Morante and Blanco (2012) and Morante *et al.* (2011), we define these elements as follows:

   - **cue**: the word (e.g. *not*), morpheme (e.g. *un-* in *un*usual) or multi-word unit (e.g. *by no means*) inherently expressing negation. The cue is the 'anchor' of textual negation, i.e. in order for an instance of textual negation to exist, there has to be a cue.
   - **scope**: all the elements whose falsity would prove negation to be false. In other words, the scope includes all those elements in a sentence whose truth value can influence the truth value of negation itself. For instance in the sentence *I am* **not** *going to school* where *I*, *am going* and *to school* are part of the negation scope, if we invert the truth value of *to school* (*to somewhere but not a school*), negation will not hold anymore. The cue is not considered part of the scope. Furthermore, it is important to notice that the scope might be discontiguous (e.g. '*I*, who was majoring in English Literature at the time, **never** *had the chance to meet him*.')
   - **event**: the element in the scope the cue directly refers to (e.g. "He is not driving a car"). Even if there might be different interpretation of what a negation event is (e.g. a hierarchy of semantic events — for instance *He is not driving a car* involves the event of 'moving', 'driving' and 'car-driving'), we consider here as event the **lexical** unit that is directly negated. Events are usually associated with the predicate negation has a scope on: verbal (as in 'He is not driving'), adjectival (as in 'He is not beautiful') and nominal (as in 'He s not a professor). Even if the cue is not included in a predicate VP, the event is considered the head of the predicate of the clause that contains the cue (e.g. [**Nobody**]$_{\text{NP}}$ [likes$_{\text{V}}$ spaghetti]$_{\text{VP}}$). Unlike previous work, when the event was defined as the minimal unit the cue directly refers to, we here consider the event as including auxiliaries, copulas or modifiers it may appear with. Finally, a negation instance might or might not contain an event (e.g. interjection as in 'Do you want to buy it? **No**, thanks' or verbal ellipsis as in 'She swims but I do **not**').[2]

   Given the three elements above, the problem of translating negation can be then redefined as the problem of correctly translated cue, event and scope. Given that these constituents are not language specific, a first advantage of decomposing negation is that all typological variations conform to a set of three basic elements. In the context of SMT, there is also the advantage of reducing negation into tangible elements at the string level. Finally, given that we have defined a set of categories to work with, we are now able to classify the errors made by a machine translation system more precisely.

4. Understanding the problem of translating negation (i.e. "What are are the errors made by translation systems?", "Do different language pairs and systems show different error patterns?"):

   - Evaluate the translation of negation manually: we conducted a manual analysis of the errors involved in translated negation for the Chinese-English and English-Korean language pair, showing that different languages are affected by different error patterns. Although beyond the set of languages explored by the HimL project, we believe that such exploratory analysis can be easily applied to European languages as well.

Based on these findings, we are planning to undertake the following steps:

[2] Although not in the scope of the present work, previous literature has also considered the *focus*, i.e. the part of the scope that is directly negated or more emphasized, as a sub-constituent of negation. Focus is the most difficult part to detect since it is the most ambiguous. For example, in the sentence 'He does not want to go to school by car' the speaker might emphasize the fact that '*He* does not want to go to school by car' or that 'He does not want to go *to school* by car' (but he wants to go somewhere else) or that 'He does not want to go to school *by car*' (but by other means of transportation).

- Plan the development of an automatic semantic evaluation for negation: the goal is to be able to automatize the manual analysis so it can be extended to any language pair or system. This is because, the sub-components of negation we are working with (cue, event and scope) are in principle language-independent. To do so, we are going to:
- Develop an automatic negation detection algorithm, to detect cue, event and scope automatically.
- Develop a way to assess the overlap between machine output and reference translation with regards to these three elements.

During this first year we also published two related workshop papers:

- F. Fancellu, B. Webber (2015), Translating Negation: Induction, Search and Model Errors, Proceedings of SSST-9, Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 21-29, Denver, Colorado, June 4, Association for Computational Linguistics.
- F. Fancellu, B. Webber (2015), Translating Negation: A Manual Error Analysis, Proceedings of Ex-Prom Workshop, NAACL-HLT, pages 1-11, Denver, Colorado, June 5, Association for Computational Linguistic

3 Task 2.3: Improving Core Fidelity of Shallow Models

The goal of Task 2.3 is to improve translation quality of shallow models (mainly phrase-based and perhaps also shallow-syntactic ones) by avoiding errors introduced *by individual translation units* (i.e. phrase pairs in the phrase-based approach). Other errors may be still introduced due to some inadequate *combination* of translation units, indicating a modelling error of the approach.

We motivated the task observing that current word alignment and phrase extraction techniques allow for systematic extraction of incorrect phrase pairs. In the example:

- English: He has | very little interest .
- Czech: Nemá | velký zájem .
- Gloss: He-does-not-have | great interest .

the "|" denotes phrase segmentation and translations as licensed by current alignment and extraction techniques. If this is a pair of training sentences, the system will learn that "Nemá = He has" (while the actual meaning of "nemá" is "he does not have") and that "very little interest = velký zájem" (while the actual meaning of "velký zájem" is "great interest").

When a new sentence is translated and the system decides to use both of these phrases, the translation will be correct, with correct negative polarity. But if only one of them is found in a source sentence, with the rest of the source sentence covered using phrases extracted from other sentences, the polarity of its translation will be reversed.

Since there is no explicit link between phrases in phrase based MT (except for the limited n-gram overlap due to the language model), we assume that the translation quality can be improved by avoiding such risky or outright erroneous phrase pairs.

3.1 Errors Fixable by Avoiding Wrong Translation Units

In the analysis of our best English-to-Czech system, Chimera (Bojar *et al.*, 2013; Bojar and Tamchyna, 2015) (see Tamchyna and Bojar (2015) for a detailed analysis of the system), we checked the quality of phrase pairs from two sources: phrases extracted using the standard phrase extraction pipeline from a parallel corpus and phrases produced by a transfer-based deep-syntactic system TectoMT, which Chimera uses in a simple system combination. We looked at the percentage of such bad phrase pairs in two settings:

- phrase pairs contained in the phrase table
- phrase pairs used in the 1-best translation

We can assume that most of the noisy phrase pairs in the phrase tables are never used in practice (they are improbable according to the data or they apply to some very uncommon source phrase). That is why we also looked at phrase pairs *actually used* in producing the 1-best translation of the WMT 13 test set.

For each of the two settings, we took a random sample of 100 phrase pairs from each source of data and had two annotators evaluate them. The basic annotation instruction was: "A phrase pair is correct if you can imagine a context where it could provide a valid translation." In other words, we are checking if a phrase pair introduces an error already on its own.

		OK	Bad	Unsure	IAA
phrase table	from Corpus	76.0%	17.5%	6.5%	78.0
	by TectoMT	66.3%	26.3%	7.4%	83.0
used	from Corpus	89.0%	7.5%	3.5%	94.0
	by TectoMT	87.5%	9.0%	3.5%	87.0

Table 4: Correctness of phrases in Chimera's phrase tables.

Annotators agree	Disputable
Negation:	
we need ≠ nemusíme (we need not)	
offer ≠ nenabízí (does not offer)	
no means ≠ prostředky (means, resources)	
Content word mistranslation:	**Things that get often dropped:**
this week in ≠ tento měsíc v (this *month* in)	she ≠ už (already)
Content word missing:	provided ≠ jsou (they-are)
town hall ≠ město (town, city)	does ≠ chce (he/she/it wants-to)
images of distant ≠ vzdálených (distant)	**Differences in person:**
cd 4 count ≠ CD 4	can we ≠ může (she/he can)
think he will ≠ , že (that)	
Incl. missing pronouns:	
him . ≠ .	
put me ≠ dát (put)	

Figure 4: Types of errors of phrase-table entries where two annotators agreed that the phrase-table entry is wrong or where they disagreed about its usability in translation.

Table 4 shows the results of the annotation. As expected, the percentage of inadmissible phrase pairs is much higher in the first setting (random samples from phrase tables), 17.5–26.3% compared to 7.5–9.0%. Most phrase pairs which contributed to the final translations were valid translations (87.5–89.0%).

The phrase table extracted from TectoMT translations was worse in both settings. However, while only two thirds of its phrase pairs were considered correct, more than 87% of the phrases actually used were admissible. This shows that the system combination in Chimera is effective and the final decoder is able to pick the correct suggestions quite successfully.

Interestingly, despite the rather vague task description, inter-annotator agreement was quite high: 80.5% on average in the first setting and 90.5% in the second one.

We extend this analysis by checking the types of errors observed within phrase-table entries. Figure 4 lists the typical examples, confirming that reversed negation is a prominent reason for marking a phrase-table entry as wrong, followed by issues with content words.

So far, our analyses started from erroneous phrase table entries, since in this task, we try to avoid them. Aside from this, it would be also useful to have the global view, to have the answer to the question: "What proportion errors in MT output can be fixed by avoiding wrong phrase table entries?" This question can be partially answered by educated guess, by looking at wrong MT outputs and considering the size of the span and other aspects of MT choices in the output.

We used the WMT 2015 manual ranking data to find bad sentences produced by our English-to-Czech MT system Chimera (Bojar and Tamchyna, 2015). WMT rankings are relative, so we had to come up with some heuristics identifying which *relative* rankings say that Chimera produced an *absolutely* bad sentence. One such example of a typical dismissive ranking is illustrated in Figure 5.

If we consider annotation screens (WMT15 "HITS") where ranks 1 or 2 are assigned to a system (there is a very good or good competitor), Chimera got rank 5 (Chimera was very bad) and rank 4 is free (Chimera was distinctly worse), we get 33 screens out of the total of 2709 screens mentioning Chimera. Admittedly, this is a small sample to get a reliable statistic but it is the easiest way how to get to outputs of a top-performing MT system that are bad on the absolute scale.

Table 5 details our analysis of errors in these sentences. In 7 cases, our heuristics failed and the sentences were actually acceptable translations (when compared with the reference). In 3 cases, the fluency of the sentences was obviously distorted, but this type of error is beyond what we expect can be fixed within individual phrase-table entries.

Unfortunately, the most frequent errors, as listed further down the list, are mostly impossible to avoid by removing wrong phrase-table entries. On the positive side, the most frequent error (participants' roles badly translated) is going to be dealt with

System	Best (1)	2	3	4	5 (Worst)
A					•
B					•
C			•		
D		•			
E	•				

Figure 5: A sample WMT manual ranking that suggests that systems A and B produced bad translations, not acceptable translations that just happened to sound worse than those from the other systems.

#	Error	Fixable in phrase table?
7	Unclear Why Bad	–
3	Fluency	No
6	Participants' Roles	No
3	Lexical Choice (Verb)	No
4	Misleading	No
2	Refl-Possesive	No
2	Verb (Missing)	Hopefully
1	Neg Lexicalized	Maybe
1	Content Missing (Subject)	Hopefully
1	Lexical Choice	No
...		

Table 5: Types of errors in bad Chimera outputs. The last column contains our impression whether such an error can be fixed by avoiding bad phrase-table entries.

in Task 2.1 in this workpackage, see Section 1.

Based on the listing, we see chances for "core fidelity" (avoiding bad phrase pairs) in two areas: (1) avoiding negation flip, and (2) ensuring content words do not get lost. We expected a larger repertoire of errors, but sadly, the data do not seem to allow for more. In this task, we will thus focus on these two types of errors.

It should be noted that (1), avoiding errors in negation, is separately dealt with in Task 2.2, see Section 2. Here in Task 2.3, we take a more direct and pragmatic approach, see Section 3.2 below.

For (2), preserving content words, we plan experiments for the next year.

3.2 Preventing Negation Flip

As thoroughly discussed in Section 2 and Appendix A, translation of negation or antonyms is a problematic issue in SMT. Despite the severity of the errors, only a few recent publications on this topic exist.

The existing solutions of this problem include linguistic analysis and hand-tailored algorithms that handle negation for a given language pair (Jang and Kim, 2015), extending corpora with original data synthetically rephrased to negated sentences (Wetzel and Bond, 2012), or the careful annotation of negation as carried out in Task 2.2 (Section 2).

The correction of negation is also handled by various post-editing systems such as Depfix (Rosa, 2014). In this task, we try to avoid errors in translating negations in a phrase-based translation system. In particular, we focus on errors caused by insertion or deletion of negation.

From experience, we know that a source word can easily be aligned with its opposite or negated form on the target side with no negation markers included in the alignment. For example, consider the sentence pair in Figure 6 which results in the extraction of an incorrect phrase pair "I have = nemám (I don't have)" as in the motivating example at the beginning of Section 3.

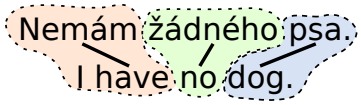

Figure 6: The typical result of automatic word alignment and phrase extraction, leading to the loss of negation.

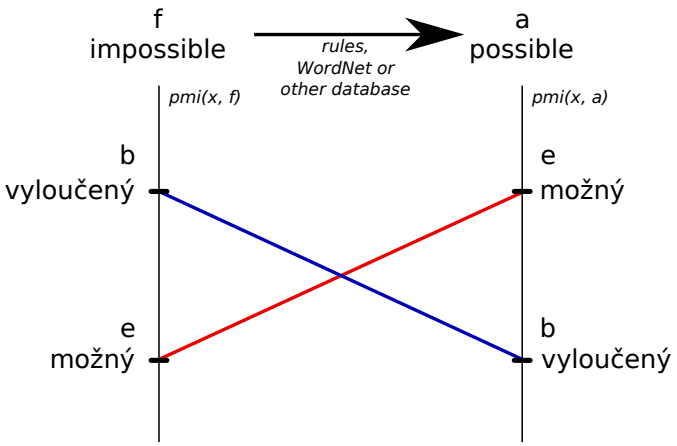

Figure 7: An illustration of a suspicious alignment pair $f - e$. The word "impossible" (f) is aligned with "možný" (e), but there exist an antonym a which includes e higher in the list of its translations (sorted by the pointwise mutual information) than a word b ("vyloučený"), which is on the other hand a better translation of f. If $f - e$ is a suspicious alignment pair then $a - b$ is very likely also a suspicious alignment pair, subject to the various thresholds.

We propose a statistical method of identifying the spots in aligned parallel data where a word is aligned to its negated form in the other language, as in the example above. A more detailed description of this method is given in Section 3.2.1.

We conduct two experiments with machine translation: (1) we filter out phrases that do not match in polarity from the phrase table in Section 3.2.2, and (2) we repair the word alignment by adding an alignment link to the negation cue if a word is aligned to with its antonym but no negation cue in Section 3.2.3.

3.2.1 Identifying Suspicious Alignments

In order to identify errors in alignments, we create a list of word-antonym pairs on the source side. WordNet (Miller, 1995) was used for this purpose. Besides WordNet, one can also obtain a list of antonyms by simple rule-based generation (such as adding the prefixes *in-* or *un-* to the word), consulted with the actual count of appearances of the generated antonym in the data.

Now, for each word f on the source side and its translation e, we estimate the pointwise mutual information $pmi(e, f)$ of e being the translation while f is on the input. We can think of the pointwise mutual information as a measure of quality of a translation. It is computed as follows:

$$pmi(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)} \tag{1}$$

We assume that in parallel data, more sentences are positive than negative and that the translation is more likely to preserve the location of negation than to move it to another element in the sentence. As a consequence, each word is more likely to be translated using the word of the same polarity.[3] This allows us to identify *suspicious alignments*: alignments of words that are very likely (cross-lingual) opposites of each other.

The alignment between a source word f and a target word e is called *suspicious* if there exists an *evidence pair* (a, b) of two words aligned to each other such that:

$$pmi(e, a) > pmi(e, f) \wedge pmi(b, f) > pmi(e, f) \tag{2}$$

In other words, we want to find an antonym a of the target word f and a word in the source language (b), such that first, e is a better translation of a than of f, and second, f is a better translation of b than of e. See the example in Figure 7.

Once we find an evidence pair for an alignment $f - e$, we consider $f - e$ suspicious in terms of preserving negation. The antonym a of f can be regarded as a correct (better) translation of e, which means that the meaning of f has to be negated (or inverted) to be the proper translation of e. Since we assume that the sentences as whole correctly preserve the meaning, we expect to

[3] Many exceptions exist; based on a quick analysis performed by a colleague of us, Martin Popel, using a half of the CzEng corpus, about 0.9% of nodes in the tectogrammatical layer of annotation (which roughly correspond to content words) do not have the same value of the negation flag. The negation is expressed somewhere else in the sentence. Prominent examples are e.g. "to není možné (it is not possible) — it is impossible", "se nepotvrdil (was not confirmed) — had proved it is not the case", "neoprodleně (immediately) — without any delay", "sotva postřehnutelný (hardly perceptible) — imperceptible".

find a negation cue somewhere in the vicinity of the word f. Section 3.2.3 describes the experiment where we add the missing alignment link between the word e and the negation cue.

In order to work with incorrect alignments only with high confidence, we apply thresholds to counts of the individual words seen in the data as well as to counts of pairs of words aligned to each other. We also apply a threshold to pointwise mutual information values computed to filter out poor evidence pairs.

Among the suspicious alignments, we select the incorrect ones by applying another threshold value to the actual difference between $pmi(e,a)$ and $pmi(e,f)$, and another threshold value to the difference between $pmi(b,f)$ and $pmi(e,f)$.

3.2.2 Phrase Table Filtering

Our initial experiments were conducted using Moses (Koehn *et al.*, 2007) and the plain phrase-based model for our WMT15 submission (Bojar and Tamchyna, 2015). This is a very competitive setup thanks to the large parallel and monolingual data.

From the phrase table, we filter out phrase pairs that contain the word *not* on the source (English) side and do not contain any negated word on the target (Czech) side. The phrase table was constructed with words represented as the pair of word form and morphological tag on the target side. Czech morphological tags include an indicator of negation, so finding the problematic pairs is very easy.

We apply the filtering after the MERT optimization, on the phrase table limited to entries needed for the translation of the test set. The weights of the model are thus kept intact and the filtering is very fast.

The original phrase table contains about 26.4M entries and the filtered one contains 26.0M entries. The filtering thus removes about 1.35% of entries. Note that the phrase table is already restricted to the test set.

The baseline model achieves BLEU score of 23.29 and the filtered model slightly improves on this, reaching BLEU of 23.36. A small manual evaluation also confirms the improvement: negation does not get lost that often.

Note that this initial experiment did not make any use of the automatically identified pairs of words and their antonyms in the target language. We will extend the filtering and remove also phrase pairs containing such bad pairs of words and missing any negation cue.

3.2.3 Fixing Alignments

One of the possible utilizations of the list of incorrect alignments is to try and fix them.

In short, we process all the word alignment pairs and if the alignment is *suspicious* as defined in Section 3.2.1 above, we ensure that a negation cue must be also linked to this word pair. Adding such a link prevents phrase extraction from extracting the wrong phrase pair.

We search for the cue in neighboring positions around both words from the alignment. If there is one found, we add an alignment between the cue and the word on the other side of the alignment. We ignore the alignment when the cue is found on both source and target windows (which is a relatively rare event). The selected window in our experiments was 3 words to the left and 1 word to the right of the current word. The cue words searched on the English side include *not, none, n't, hardly, little,* etc. On the Czech side, every word whose tag indicates it is negated is considered a potential cue word.

When fixing alignments in the training corpus of 52,557,627 sentence pairs, only 30,418 were fixed. This amounts to only about 0.06% of repairs. The main reason lies probably in the fact that our search for negation cue is very limited and we fail to find the cue for many suspicious alignments.

The translation table extracted using these fixed alignments contained 34,242 fewer extracted phrases than the translation table extracted following the original alignments. This is a very little change, only 0.005% of entries, since here we work with the full phrase table of 671M entries, not just the version limited to a given test set.

As a result, the BLEU score did not change and even the manual annotation did not reveal any improvement.

3.3 Conclusion

In Task 2.3, we tried to improve core fidelity of the phrase-based model by avoiding phrase pairs that are very likely to introduce a translation error. A detailed analysis revealed that there are fewer opportunities for such corrections than anticipated, but they are still worthwhile.

We started by avoiding negation flip, proposed a method for identifying suspicious word alignments and got promising results in our initial experiments with phrase-table filtering. The suspicious alignments have yet to show their utility.

In the coming year, we will continue the work on avoiding negation flip and we will also design methods for the other good opportunity: ensuring that content words are not lost in translation.

4 Task 2.4: Employing High-Quality Large-Scale Dictionaries

Task 2.4 is planned to start in year 2.

Conclusion

This deliverable D2.1 of the HimL project documents that the work in WP2 Semantically Motivated MT runs smoothly, without any deviations from the plan.

A Challenges in translating textual negation

Abstract

Amongst a wide range of linguistic phenomena Statistical Machine Translation has addressed, negation is still one that has not yet received adequate treatment. While techniques have been proposed and implemented for improving translation performance on negation, they all follow from the developers' beliefs about why performance is worse. These beliefs, however, have never been validated by a thorough understanding of what negation is and what the errors involved in translating it are.
On the other hand, the present work starts by considering *what is meant by negation* and *what makes a good translation of a negative sentence*. We first consider breaking down negation into its sub-constituents (cue, event and scope) to allow efficient manipulation of a semantic phenomenon in a primarily string-based task, such as SMT. Focusing on these three elements, we present the plan for a semantic evaluation potentially applicable to any language pair or system. We will discuss the negation-related elements that need to be evaluated, how it is possible to detect them automatically across different languages and ways their translation can be best evaluated.
We hope that such investigation will guide future work on improving the translation of negative sentences, customisable for different language pairs and machine translation systems.

A.1 Introduction

A.1.1 The problem

Since its early years, Statistical Machine Translation (SMT) has addressed the problem of modelling specific linguistic phenomena whose realisation varies across languages. For instance, phenomena such as pronoun or tense translation (Hardmeier and Federico, 2010; Gong *et al.*, 2012) are problematic because they involve a mapping operation not only on strings or constituents but also on a set of linguistic information that are sometimes non-local, into a language that might encode this information differently or not at all.
Negation also falls into this category. Although it is widely accepted that negation is a semantic universal i.e. every language has a way to reverse the truth value of a statement, the way each language realises it differs greatly, making it difficult for tasks involving cross-lingual prediction, such as Machine Translation.
In the NLP community, the importance of negation has been acknowledged in fields other than SMT. Automatic detection of negation, for instance, is an important part of Biomedical information retrieval where detecting the truth value around one or more named entities (e.g. "**Mr. Lee** shows no signs of **MERS virus**") allows quick information extraction that can be subsequently analysed and stored. The same holds in an automatic translation scenario, where preserving the truth value of a statement can have serious implications on the quality of the translation, especially if applied to the medical domain. However, despite the potential gain in ensuring that negation related information is correctly translated, only limited efforts have been made in SMT. Moreover, most of these efforts either briefly consider negation in the context of other problems - such as reordering (Collins *et al.*, 2005) or formulate a hypothesis about what might go wrong when translating negation without any previous error analysis or thorough understanding of the phenomenon itself. The gaps in previous literature are then both theoretical and practical. In contrast, we want to better understand what it is meant by negation and which of its aspects are to be considered from a cross-lingual, computational perspective. Furthermore, we want to know what errors an SMT system makes in translating negation, so as to avoid tentative guesses. Our challenging goal is to have an approach to translating negation that fits into the SMT pipeline and can be used for many, if not all, language pairs.

A.1.2 The scope of the project

Before introducing the approach taken in this work, it is important to clarify its scope. This work takes into consideration the translation of what we define as overt textual negation — i.e. negation that is explicitly marked by means of either lexical, morphological or syntactic processes. This is different from translating the sentiment of a sentence, which is usually defined by the presence and combination of certain words with either a positive, negative or neutral connotation, which may not be overt negation items at all. This work also does not deal directly with certain lexical elements that do not contain an explicit negation marker but have an inherent negative meaning (e.g. failed to X → did not X). We will however take into consideration the possibility that an item overtly negated can be translated into its antonym (e.g. not expensive → cheap, not succeed → fail). An important aspect of this work revolves around understanding what is negation and how this knowledge can be used to enhance an SMT system. The main issue arises when we consider that the entire SMT pipeline is based on strings or a combination of those whilst negation is a rather abstract semantic phenomenon. The present work tries to bridge this gap by breaking negation down into its sub-constituents (cue, event and scope) and ground them at a string level. An immediate advantage of this analysis is that we can keep track of each of the negation elements during the translation process so as to guide their translation and interaction.

A.1.3 First year milestones

The first year milestones we will discuss in the present report are as follows:

1. Getting acquainted with previous work on the topic
2. Understanding where the gaps in previous work are
3. Understanding negation, with particular emphasis of its translation

4. Understanding the problem of translating negation (i.e. "What are are the errors made by translation systems?", "Do different language pairs and systems?")
   - Evaluate the translation of negation manually
   - Plan the development of an automatic semantic evaluation for negation

   Although the languages here considered are not the ones the project is directly involved with, we believe the conclusion drawn from this analysis are easily applicable to other language pairs.

During the first year we also published two related workshop papers:

- F. Fancellu, B. Webber (2015), Translating Negation: Induction, Search and Model Errors, Proceedings of SSST-9, Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 21-29, Denver, Colorado, June 4, Association for Computational Linguistics
- F. Fancellu, B. Webber (2015), Translating Negation: A Manual Error Analysis, Proceedings of Ex-Prom Workshop, NAACL-HLT, pages 1-11, Denver, Colorado, June 5, Association for Computational Linguistic

A.2 Background

A.2.1 Negation in SMT

The starting point of this work is to verify whether negation is indeed a problem in SMT and if so, why do we consider it an unsolved problem.

Previous work have pointed out the problem of translating negation for different language pairs: in a Japanese-to-English translation task, (Wetzel and Bond, 2012) show that the translation system performs worse in terms of BLEU score on a sub-set of only negative sentences compared to one of only positive sentences; (Fancellu and Webber, 2014) show that the same holds from a Chinese-to-English hierarchical phrase-based system; (Popovic and Arcan, 2015) point out the problem of translating from English into South Slavic languages, where SMT systems fails in rendering the double negation structure correctly; finally in (Bojar *et al.*, 2013), the problem of correctly rendering negation is observed when translating into Czech.

In general, previous work that have tackled the problem of translating negation can be grouped according to what they hypothesised the cause of this problem is as follows:

- **Structural mismatch between source and target language**: (Collins *et al.*, 2005) and (Li *et al.*, 2009) see the problem of translating negation as 'positional', given that different languages have different ways of placing the negative marker with respect to other elements in the sentence. Both works place negation in the context of clause restructuring via pre-processing of the source sentence.
  In (Collins *et al.*, 2005), the German source side is pre-processed both before training and decoding so as to match the word order of the target language. More specifically, negation is moved to a position that can be easily captured by phrase-based models. For instance, when the German verbal phrase [[konnten]$_{mod}$ [einreichen es]$_{V}$ [nicht]$_{neg}$]$_{VP}$ is restructured, the negative marker is moved back one position to match the English word order [[could]$_{mod}$ [not]$_{neg}$ [hand it in]$_{VP}$]$_{VP}$.
  (Li *et al.*, 2009) take into consideration translating from Chinese to Korean in the context of phrase-based SMT. Given that the two languages exhibit different word orders (SVO and SOV respectively), they propose a rule-based pre-reordering approach to match the syntax of the target language. Part of their work explicitly addresses negation where the negation marker can be separated from the verbal head in Chinese whist cannot in Korean. For instance, in [1~3], a propositional phrase can intervene between the negative marker and the verbal head in Chinese, while the corresponding Korean sentence does not allow such gap.

  (1) *Chinese*: **bù**$_{neg}$ yīnggāi$_{md}$ [yǐ guǎnlǐyuán shēnfén]$_{PP}$ yùnxíng$_{V}$
  not should as administrator credential run
  'it should not be run as administrator'

  (2) *Korean*: [gwanlija gwonhan-eulo]$_{\text{PP}}$ [silhaenghamyeon$_{\text{V}}$ **an$_{\text{neg}}$** doeda]$_{\text{VP}}$

  (3) *[silhaenghamyeon$_{\text{V}}$ [gwanlija gwonhan-eulo]$_{\text{PP}}$ **an$_{\text{neg}}$** doeda]$_{\text{VP}}$

  However, the work seems to focus only on bringing the negation marker close to the verbal head but not inverting its order to a post-verbal position.
  Both works report an improvement on translating quality, measured in BLEU scores.
- **Lack of negative training data**: (Wetzel and Bond, 2012) address the problem of the lack of sufficient negative training data as the reason why translation performance of negative sentences (as measured by BLEU) is worse than positive sentences; in the parallel corpora considered, in fact, only ~20% of the sentences contain one or more instances of textual negation. To solve this problem, the training set is enriched with negative paraphrases of positive sentences; these are first transformed into MRS (Minimal Recursion Semantics, (Copestake *et al.*, 2005))-based representation and a negative 'handle' is added onto the main verb (i.e. the verbal handle at the top at the MRS graph). This process is applied to both the English and the Japanese sentence. Improvement over the baseline on a sub-set containing negative sentences only (+1.63) is observed only in the case when negation paraphrases are also added to the language model.
  A manual inspection of those sentences highlights how uninformative n-gram overlap-based automatic metrics are, when evaluating the translation of negative sentences (here, taken to be whether the negation marker is correctly reproduced

in the target side when present on the source). Although this approach leads to 33 sentences being translated correctly when the baseline does not, at the same time the enriched model fails to translate negation correctly in a similar number of sentences whereas the baseline does. Moreover, the number of what is referred as 'critical negation-related errors' is reduced merely from 69 to 66.

- **Lack of any negation related information**: Another line of work poses the question of whether the problem lies instead in the model not containing specific information on how negation is expected to be translated. Unlike (Wetzel and Bond, 2012), these work assume that there is sufficient training data to learn how negation is translated but the model lacks any guidance on how to use them. (Baker *et al.*, 2012) and (Fancellu and Webber, 2014) try at first to break down negation into sub-components whose presence and interaction the model can then assess.
  (Baker *et al.*, 2012) take into consideration enhancing both source and target side syntactic trees with negation related information in the context of an Urdu-to-English tree-to-tree syntactic translation task. This is done by adding the notion of *trigger* and *target*, respectively the negative marker and the entity it directly refers to, on both the direct pre-terminal associated to each of these elements and their common ancestor. The intuition here is that, when during decoding, we are scoring the likelihood of generating a negated VP, we would expect to give a higher score to those derivations that also contain a trigger and a target node. Results shows little, although statistically significant, improvement over the baseline when those rules are enhanced. However, given that the scope of this work includes modality as well and given that no concrete manual analysis is carried out, it is not clear where this improvement comes from.
  On the other hand, (Fancellu and Webber, 2014) try to answer the question of whether it is the model being responsible for an observed worse translation performance on negative sentences by investigating n-best lists output after decoding. If these contain a better translation of a negative sentence than the 1-best, it can be concluded that the model can potentially generate negative sentences but need guidance in promoting those.
  The lookup for a better translation is done by decomposing negation into *cue* (same as the *trigger* above), *event* (similar to the *target*) and *scope* (see §2.2 for more detail). In an oracle experiment setting, these elements are approximated in the n-best list hypothesis and in the reference translation using a dependency parse and then compared. The sentence whose negation elements are the most similar to the reference translation is brought to the top and automatic evaluation (re)performed. The oracle leads to a considerable boost of more than 4 points BLEU showing that the model is able to output a better translation of negation than the baseline. A second experiment tries to use lexical translation probabilities to score these elements by only relying on the source sentence, whose negation elements are again approximated in a dependency parse structure.

Although this works differ in the reasons they give on why SMT systems perform worse when translating negative sentences, common shortcomings are worth highlighting:

1. Negation is only considered in a few works and it is often treated as a side problem. (Collins *et al.*, 2005) and (Li *et al.*, 2009) consider it among other linguistic phenomena in the bigger picture of clause restructuring and re-ordering; (Baker *et al.*, 2012) take it into consideration only when it is associated with modality; finally (Popovic and Arcan, 2015) and (Bojar *et al.*, 2013) take negation into account only amongst other types of errors the SMT system produces.
2. Even when negation is the focus of a study, the scope is narrow and only one hypothesis or language pair is considered.
3. None of the previous studies has conducted an extensive error analysis on the phenomenon so it is not clear what exactly are the kind of errors produced by an SMT system and what those depend on. This is connected to (2), where an hypothesis is formulated without any thorough inspection of the data.

A.2.2 Decomposing negation

Analysing the errors involved in translating negation leads us to the question of what is meant by negation and what are we assessing the translation of. Previous work that have specifically addressed the problem of translating negation ((Wetzel and Bond, 2012) and (Bojar *et al.*, 2013) in particular) have cast the problem of translating negation as a binary outcome of whether a single element, the negation marker, has been translated or dropped during translation. Reducing it to a presence vs. absence problem is however an over-simplification. The negative marker interacts in fact with certain elements in the sentence whilst some others are outside its scope; this implies that if during translation, certain elements that are not under the scope of negation are brought inside it or viceversa, the overall rendering of the negation phenomenon is not correct. For this reason, the present work starts from the assumption that negation is a structural phenomenon, composed of a finite set of elements that interact with each other. Following (Blanco and Moldovan, 2011), (Morante and Blanco, 2012) and (Morante *et al.*, 2011), we define these elements as follows:

- **cue**: the word (e.g. *not*), morpheme(e.g. *un-* in *un*usual) or multi-word unit (e.g. *by no means*) inherently expressing negation. The cue is the 'anchor' of textual negation, i.e. in order for an instance of textual negation to exist, there has to be a cue.
- **scope**: all the elements whose falsity would prove negation to be false. In other words, the scope includes all those elements in a sentence whose truth value can influence the truth value of negation itself. For instance in the sentence *I am* **not** *going to school* where *I*, *am going* and *to school* are part of the negation scope, if we invert the truth value of *to school* (*to somewhere but not a school*), negation will not hold anymore. The cue is not considered part of the scope. Furthermore, it is importance to notice that the scope might be discontinuous (e.g. '*I*, who was majoring in English Literature at the time, **never** *had the chance to meet him*.')

- **event**: the element in the scope the cue directly refers to (e.g. "He is not driving a car"). Even if there might be different interpretation of what a negation event is (e.g. a hierarchy of semantic events — for instance *He is not driving a car* involves the event of 'moving', 'driving' and 'car-driving'), we consider here as event the **lexical** unit that is directly negated. Events are usually associated with the predicate negation has a scope on: verbal (as in 'He is not driving'), adjectival (as in 'He is not beautiful') and nominal (as in 'He is not a professor). Even if the cue is not included in a predicate VP, the event is considered the head of the predicate of the clause that contains the cue (e.g. [**Nobody**]$_{\text{NP}}$ [likes$_{\text{V}}$ spaghetti]$_{\text{VP}}$). Unlike previous work, when the event was defined as the minimal unit the cue directly refers to, we here consider the event as including auxiliaries, copulas or modifiers it may appear with. Finally, a negation instance might or might not contain an event (e.g. interjection as in 'Do you want to buy it? **No**, thanks' or verbal ellipsis as in 'She swims but I do **not**').[4]

Given the three elements above, the problem of translating negation can be then redefined as the problem of correctly translated cue, event and scope. Given that these constituents are not language specific, a first advantage of decomposing negation is that all typological variations conform to a set of three basic elements. In the context of SMT, there is also the advantage of reducing negation into tangible elements at the string level. Finally, given that we have defined a set of categories to work with, we are now able to classify the errors made by a machine translation system more precisely.

A.3 Error analysis

In §2, we have observed that techniques proposed and implemented for improving translation performance on negation have simply followed from the developers' beliefs about why performance is worse. These beliefs, however, have never been validated by an error analysis of the translation output. The present work takes instead an empirical approach towards understanding why negation is a problem in SMT, by first investigating *what kind* of errors are involved in translating negation and showing that tailoring to a semantic task, string-based error categories standardly used to evaluate the quality of the machine translation output, allows us to cover the wide range of errors occurring while translating negative sentences.

The analysis is the same as we attempted in (Fancellu and Webber, 2015): we first manually annotate both source and hypothesis translation for the sub-constituents of negation introduced in §2.2 (cue, event and scope) and then use a precision-recall based metric to quantify the amount of correct translations for each element. The reason why we choose to evaluate the hypothesis against the source and not the reference translation is because there are potentially different ways the same negation structure can be rendered; if hypothesis and reference translations differ in the realisation of negation but they are equally correct, it would be difficult to quantify the errors using a string-matching, precision-recall based metric.

We apply this analysis to both a Chinese-to-English (from (Fancellu and Webber, 2015)) and an English-to-Korean translation output. We chose these two language pairs to investigate whether the similarity in realizing negation between two languages has an effect in translating textual negation.

Chinese and English show very similar patterns in expressing negation: they are both SVO languages, where the most frequent cues are pre-verbal (see (Blanco and Moldovan, 2011) and (Fancellu, 2013)); negation on existentials and copula is expressed by negating a verb with a separate negation cue; both show morphological negation on adjectives.

On the other hand, translating from English to Korean, involves the problem of translating into a morphologically rich language. English is an SVO language where negation on verbs is only expressed syntactically, whereas Korean exhibits a SOV order where negation can be either syntactical and morphological. Moreover, unlike English, in Korean there exists a separate verb form for negative existential and copula. Finally, there is a data sparsity issue when dealing with this language pair.

A.3.1 Annotation of negation

The first task is to annotate *cue*, *event* and *scope* in both source and hypothesis translation. The annotations are carried following the guidelines released during the *SEM 2012 shared task for automatic detection of negation (Morante *et al.*, 2011). It is however worth noting that while these guidelines were released with the goal in mind of automatically extracting information from text, with a particular emphasis on factuality, the present work focuses on translation, where each negation instance is taken into consideration as potential source of error. This leads to some differences in the annotation process, especially in the case of the *event*:

1. While the original guidelines do not annotate the presence of an event when this is non-factual, such as in conditional clauses ('*if he doesn't come*, I will blame you'), the demands of translation require it to be annotated.
2. While the original guidelines do not include modals or auxiliaries in the event annotation (in order to minimise the number of annotated elements), getting these elements correct in translation is needed to distinguish a correct vs. a partially correct event. In the case of resultative constructions (e.g. *fù bù qǐ* lit. 'pay not lift-RES.', 'could not pay, can not afford') we considered the resultative particle as part of the *event*.
3. For the same reason as (2), the event in a nominal predicate includes all its modifiers.

With respect to scope, the current work makes a simple approximation: scope is often discontinuous, with multiple semantic units whose translations might impact the overall translation of the scope differently. To facilitate error analysis we approximate the scope in terms of its constituent semantic roles, here taken to be underspecified PropBank-like semantic arguments. In doing

[4] Although not in the scope of the present work, previous literature have also considered the *focus*, i.e. the part of the scope that is directly negated or more emphasized, as a sub-constituent of negation. Focus is the most difficult part to detect since it is the most ambiguous. For example, in the sentence 'He does not want to go to school by car' the speaker might emphasize the fact that '*He* does not want to go to school by car' or that 'He does not want to go *to school* by car' (but he wants to go somewhere else) or that 'He does not want to go to school *by car*' (but by other means of transportation).

so, we consider the scope as the *semantic domain* of negation, where the constituent elements are expected to remain in its boundaries and to preserve their semantic role (or take an equivalent one) during translation.

The annotation process is carried out in those sentence pairs containing overt negation on the source. A second pass is then performed to spot those pairs with overt negation only on target side, which might signal potential insertion errors (see next section).

Example (4) illustrates our annotation scheme over the first instance of *bù* (not) in a Chinese source sentence.

(4) [*wǒmen*]$_{scope.role}$ *bù*$_{cue}$ *páichú*$_{event}$ [*qízhōng yǒu dǎn xīn de huì lái zhǔdòng jiāodài*]$_{scope.role}$, *dàn páo de qǐ bù gēng duōme*?
We not exclude amidst there is worried of can come voluntarily confess , but run RES not even more Q

Ref: [We]$_{scope.role}$ do not$_{cue}$ [rule out]$_{event}$ [the possibility that some timid ones might come out and voluntarily confess]$_{scope.role}$, but would n't many more just run away?

As shown in (4), the scope around the first main clause can be split into two arguments - a subject and an object - around the verb *páichú*(rule out) so error analysis can be carried on each individually.

While annotating negation we also distinguish between functional and non-functional negation, the latter which we do not annotate. This is the case of the second instance of *bù*/*not* in (4) where the non-functional cue is just part of the question; other examples of non-functional negation can be question tags (e.g. *It is cold,* ***isn't it****?*) and fixed expression containing a negation cue but having a positive meaning (e.g. *Tāmen* ***bùdébù*** *qù cānjiā tā de hūnlǐ*, 'They had **no** choice but to attend her wedding' = They attended the wedding).

A.3.2 Quantifying the errors

A subsequent task is to define categories that are able to cover potential errors in translating negation. Our analysis aims at applying a small set of string-based operations traditionally used in SMT to the aforementioned elements of negation. We take into account three main operations out of the ones first introduced by (Vilar *et al.*, 2006) and apply them to each of the three elements of negation for a total of 9 main conditions:

- **Deletion**: one of the three sub-constituents of negation is present in the source Chinese sentence but not in the machine output. This corresponds to the *missing words* category in (Vilar *et al.*, 2006).
- **Reordering**: whether the element has been moved outside its scope. Since some semantic elements can also move inside the scope and *erroneously* take a role which they did not have in the original source sentence, we define the former as *out-of-scope* reordering error and the latter as *same scope* reordering error. The reordering category represents an adaptation of the original *word order* category.
- **Insertion**: the negation element is not present in the source sentence but has been inserted in the machine output. Insertion is defined here as a negation element that is aligned to a source phrase that is not a negation element. As for a scope role or the event, this is the case of phrases that were not inside the negation scope in the source but are in the target. In the case of the cue, it implies that a new negation instance has been created where there were none in the source. This resembles the *extra words* sub-category in the *incorrect words* class.

Since we are not concerned with errors regarding style, punctuation or unknown words, other operations were left aside.

For a better understanding at *when* during the translation process and *why* the error occurs, we also investigated the trace of rules used to build the 1-best machine output. This is particularly useful in the case of deletion: this may occur because a certain word or phrase has not been seen during training (*out-of-vocabulary items* - OOVs) and the system is therefore unable to translate them.

After the elements of negation have been annotated in both the source sentences and machine outputs, we use the same heuristic as (H)MEANT (Lo and Wu, 2011) to decide whether a unit is translated correctly.

MEANT and its human counterpart HMEANT are semantic-oriented SMT evaluation metrics where the overlap between the semantic roles around a given predicate between hypothesis and reference translation is assessed and quantified via an F1 measure. (H)MEANT works as follows: first, semantic frames are annotated in both the hypothesis output and the reference sentence where semantic roles are associated to the frame a certain predicate is the head of; afterwards, semantic frames in the reference are aligned to the ones in the MT output. If the predicate of a reference frame is found in the hypothesis, the metric then proceeds to quantify the amount of overlap between the label and the content of the semantic roles. The metric also allows for partial matches of a semantic role where the core semantics is conveyed. If a reference predicate is not found in the hypothesis, it is considered a miss, no matter the amount of overlap between the roles associated to it.

The present work uses the same intuition but with some minor modifications. First, we score cue and scope overlap even when there is no match between the event (which often is the head of the predicate) of the hypothesis and in the reference. This is because we are assessing each category separately. Following HMEANT, we consider synonyms of the source negation to be *correct* translations since they are taken to convey the same meaning. This also includes those elements that are negated in the source but are rendered in the machine output by means of a lexical element inherently expressing negation (e.g. *fails*) or by paraphrase into positive (e.g. *bù tóng*, lit. 'not similar' → different). Translated elements that do not contain errors which impact the overall meaning are considered as *partially correct*. In the case of the event, this might be related to tense agreement or wrong modality, whilst in the case of the scope it is usually related to the fact that secondary elements are not translated correctly but the overall meaning is still preserved.

As in HMEANT, we compute precision, recall and F_1 measure using the following formulae where $e \in \mathrm{E} = \{\text{cue,event,filler}\}$. Normalisation is done at the text level, where we take into consideration the total number of each of these elements in the whole test set. Unlike HMEANT, we do not normalise the number of correct fillers by the number of total fillers in the predicate.

$$P = \frac{(\sum e_{correct} + 0.5 * \sum e_{partial})}{\sum e_{hyp}}$$

$$R = \frac{(\sum e_{correct} + 0.5 * \sum e_{partial})}{\sum e_{src}}$$

$$F_1 = 2 * \frac{P * R}{P + R}$$

A.3.3 System

For the Chinese-to-English pair, we deployed the hierarchical phrase based system submitted by the University of Edinburgh for the NIST12 MT evaluation campaign. The system was trained on approximately 2.1 million length-filtered segments in the news domain, with 44678806 tokens on the source and 50452704 on the target, with MGIZA++ (Gao and Vogel, 2008) used for alignment. The system was tuned using MERT (Minimal Error Rate Training, (Och, 2003)) on the NIST06 set.

Two different test sets were considered to assess differences that might be associated with genre: the NIST MT08 test set, containing data from the newswire domain and the IWSLT14 tst2012 test set, containing transcriptions of TED talks. We hypothesise that the difference in genre can influence the kinds of negation related error occurring during translation: as a collection of planned spoken 'persuasive' talks, we expect the IWSLT'14 test set to contain shorter sentences, and on average, more instances of negation. On the contrary, we expect the NIST MT08, where data are from the written language domain, to contain longer sentences and fewer instances of negation. Out of the 1397 segments in the IWSLT2014 set and the 1357 segments in the NIST MT08 set, 250 sentences for each set were randomly chosen to carry out the manual evaluation. Randomisation means that we do not control for the presence of negation in the sample considered.

To quantify the errors in the English-to-Korean translation task, we train a hierarchical phrase based model on approximately 213000 segments, with 2848450 tokens on the source and 2440867 on the target side. The development data and the test data contain 1500 and 2000 sentences respectively. All data was created by combining together the KAIST parallel corpora[5] with manually translated TED talks (Cettolo *et al.*, 2012). No lemmatisation or other processes were applied to the target side; future work might address this point. As with the Chinese-to-English task, we isolated 250 random sentences to carry out the manual evaluation.

A.3.4 Results: Chinese-to-English: NIST MT08

The results of the manual evaluation for the NIST MT08 test set are reported in Table 6. It can be easily seen that getting the cue right is easier than translating event and scope correctly. The cue is in fact usually a one-word unit and related errors concern almost entirely whether the system has deleted it during translation or not. Event and scope instead are usually multi-word units whose correctness also depends on whether they interact correctly with the other negation elements.

In those cases where the cues were deleted during translation, the trace shows that they were all caused by a rule application that does *not* contain negation on the English right hand side. Also worth noticing is that, in these cases, the negation cue in the source side is lexically linked to the event ('***bù**shǎo*' , 'not few, many') or lexically embedded in it (e.g. '*dé **bù**dao*, 'cannot obtain'). No cases of cues being deleted were found where the cue is a distinct unit. Also, no cases were found of cues being deleted because of not being seen during training (*out-of-vocabulary items*).

Other cue-related errors involve the cue being re-ordered with respect to scope. In one case, cue reordering happens within the same scope, where the cue is moved from the main clause to the subordinate. In three other cases, the cue is instead translated outside its source scope and attached to a different event. The two cases are exemplified in Ex. (5) and Ex. (6) respectively.

(5) [*tā*]$_{\text{sem. role}}$ *cóngbù*$_{\text{cue}}$ [*yīnwèi wǒ gěi tā tí guò yìjìan*]$_{\text{sem. role}}$ *ér* [*dùi wǒ*]$_{\text{sem. role}}$ *huài yǒu*$_{\text{event}}$ [*pìanjìan*]$_{\text{sem. role}}$ [...]
He never because I to him raise ASP opinion so to I have bias

Ref: He **never** showed any bias against me because i 'd complained to him [...]

Hyp: he **never** mentioned to him because my opinions and i have bias against china [...]

(6) [...] *jiù huì rènwéi bù*$_{\text{cue}}$ *cúnzài*$_{\text{event}}$
[...] then can think not exist

Ref: [...] people would think [that they do **not** exist]$_{\text{sub}}$

Hyp: [...] do **not** think [there is a]$_{\text{sub}}$

In (5) the cue **never** is moved from 'have bias' to the translation of the verb in the subordinate, while the opposite happens in (6) where from the subordinate the cue is moved to the main sentence.

As for the translation of events, a trend similar to the translation of cues can be observed, although the percentage of deletions is higher. The trace shows that in 3 out of 11 cases, deletion is caused by an OOV item. The remaining cases resemble the case of the cue, insofar as no rule contains the target side event. Another problem arising with events is that some scope portions in the source might have erroneously become events in the machine output and vice versa; we found 3 events on the source becoming part of the scope in the target and 7 scope portions on the source becoming events in the machine output, as shown in (4).

[5] Freely available at http://semanticweb.kaist.ac.kr/home/index.php/Home

NIST MT08 test set - 250 sentences			
Average Sentence Length	28		
Number of negated sentences	54	21.6%	
Cue per sentence ratio		1.22	
	Src	Hyp	
Cues	66	57	
Events	66	57	
Scope roles	98	80	
	R%	*P%*	F_1
Correct cues	87.87 (58/66)	92.98 (53/57)	**90.35**
Correct events	51.51 (34/66)	50.88 (29/57)	
+ Partial events	57.63 (4 + **8**/66)	57.9 (29 + **8**/57)	**57.74**
Correct scope roles	48.97 (48/98)	56.25 (45/80)	
+ Partial scope roles	58.16 (48 + **9**/98)	67.5 (45 + **9**/80)	**62.48**
Deleted cues	6 (4/66)		
Deleted events	16.6 (11/66)		
Deleted scope roles	9.18 (9/98)		
Inserted scope roles		2.5 (2/80)	
Reordered cues *same scope*	1.5 (1/66)	1.75 (1/57)	
Reordered cues *out of scope*	4.5 (3/66)		
Reordered events *same scope*	4.5 (3/66)	12.2 (7/57)	
Reordered events *out of scope*	1.5 (1/66)		
Reordered scope roles *same scope*	8.16 (8/98)	6.25 (5/80)	
Reordered scope roles *out of scope*	21.41 (21/98)		

Table 6: Results from the error analysis of the 250 sentences randomly extracted from the NIST MT08 test set.

(7) *zhè yīge jiēduàn de biǎoxiàn shì* [*duǎnqī xiāoguǒ*]$_{\text{sem. role}}$ *bù* *dà*$_{\text{cue + event}}$ [...]
This one stage of show is short-term result not big [...]
Ref: what this stage brings forward is : modest success in the short-term [...]
Hyp: this is a stage performance are not$_{\text{cue}}$ [short-term effect]$_{\text{event}}$

The fact that most reordering errors are scope-related is connected to the lack of semantic-related information during the translation process, a common problem in machine translation systems. Since there is no explicit guidance as to which events the roles in the scope should be attached to and in what order, *in-scope* and *out-of-scope* problems are to be expected.

Around 10% of scope-related errors were caused by deletion. An investigation of the trace shows that in all 9 cases, the system has knowledge of the source words in the rule table but has applied a rule that does not translate a scope portions on the target side.

Finally we notice that 2 of the incorrect scope roles in the hypothesis were due to the *insertion* of scope portions not present in the source side. The trace shows that this kind of error is generated by rules that contain on the right hand side extra material not related to the source side. We hypothesised that these rules might have been created during training where English words that did not correspond to any Chinese source words were arbitrarily added to neighbouring phrases. For instance, in (8) a rule that translates *yǐzhìyú* ('to the extent of') into 'to the extent of *they*' is used, adding a portion to following negation scope.

(8) [...] *yǐzhìyú* *wúfǎ* *yú* *oū zhou méngguó zhèngcháng zhǎnkāi hézuò*
[...] to the extent not possible with Europe union normally open cooperation
Ref: [...] even made it is impossible to carry out cooperation with their European allies as normal .
Hyp: [...] to the extent that [they]$_{\text{sem. role}}$ are unable to conduct normal with its european allies cooperation

Beside quantifying the errors for each of the negation element considered, we also analysed the kind of negation instances in the source sentences and whether there is a 1-to-1 correspondence in the way they are rendered in the reference translations.

Out of the 66 negation instances found in the NIST '08 test set, only 8 are instances of non-VP negation and two cases where the adjectival VP in the source is rendered as a non-VP negation in the target. As for the sentential environment where negation is realised (subordinate vs. main clause), we found out that there is a 1-to-1 correspondence between source and reference; this might be due to the fact that syntactically English and Chinese are very similar.

A.3.5 Results: Chinese-to-English: IWSLT '14 Tst2012 TED Talks

Results for the TED talks test set are reported in Table 7. It can be observed that results on all three categories are better than the NIST08 test set, in particular for the F_1 measure of correct events and scope. A reduction in the percentage of reordered scope roles on the overall number translation errors might be connected to shorter sentences in TED talks then in the NIST test set and with less chance of a long range reordering.

We can also observe that genre has an effect on negation cues; despite sentences being shorter, we found more negative instances in the TED talks.

As for the errors in the NIST08 test set, we analysed the trace output after the completion of the translation process to see whether deletions were caused by incorrect rule application or by the presence of OOV items not seen during training. Out of 7 cases of cue deletion, 5 of event deletion and 4 of scope role deletion, only one was caused by the presence of an OOV

vocabulary item in the source. However, as shown in (6), the OOV error is generated by a wrong segmentation of two elements in the source, *bùzhī* and *zěnme*, which end up being collapsed in a single word unit.

(9) *bùzhīzěnme* *yòng wǒmen bù néng* *wánquán lǐjiě* *de fāngshi* [...]
do not know how use we not be able completely understand of method [...]
Ref: ways we cannot fully understand that we don't know how to use [...]
Hyp: was converted to the way we cannot fully understand [..]

This seem to exclude OOV items as a problem in translating negation for the present system and what we are left with is a problem of negative elements not correctly reproduced on the target side of the rules.

Finally, we have found two cases of insertion, one cue and the other event related. Overall, cases of insertion are rare and do not constitute a real problem for the system here considered. In general, as for *event* and *scope*, a rule application that does not contain one of these two elements on the Chinese left hand side but inserts it in the English right hand side might be just fortuitous. As in the case of (8), it might have been that a rule containing extra material was preferred because a better fit in that specific context (given that a LM part of the scoring function of a SMT system). Insertion of the *cue* deserves instead a better investigation. The results shows that deletion is sometimes associated with rules whose Chinese (left-hand) side contains a cue whilst the English side does not. This is most certainly caused by the training process where rules are extracted according to what portion of the source Chinese sentence is aligned to what portion in the target English sentence. If an Chinese sentence contains negation but the English does not, a rule learnt from that pair might learn that a negation cue corresponds to something positive. This should theoretically happen the other way around and if so, the application of these rules should lead to insertion. Further analysis of the rule table and the sentences used in training might clarify this point.[6]

IWSLT14 tst2012 TED talks - 250 sentences			
Average Sentence Length	18		
Number of negated sentences	61	24.4%	
Cue per sentence ratio		1.13	
	Src	Hyp	
Cues	69	54	
Events	69	52	
Scope roles	103	83	
	R%	*P%*	F_1
Correct cues	88.4 (61/69)	98 (53/54)	**92.95**
Correct events	69.56 (48/69)	76.92 (40/52)	
+ Partial events	71.73 (48 + **3**/69)	79.8 (40 + **3**/52)	**75.55**
Correct scope roles	62 (64/103)	77 (64/83)	
+ Partial scope roles	63.59 (64 + **3**/103)	78.9 (64 + **3** /83)	**70.42**
Deleted cues	10.14 (7/69)		
Deleted events	7.2 (5/69)		
Deleted scope roles	3.8 (4/103)		
Inserted cue		1.8 (1/54)	
Inserted scope roles		1.2 (1/83)	
Reordered events *same scope*	7.2 (5/69)	1.9 (1/52)	
Reordered events *out of scope*	5.7 (4/69)		
Reordered scope roles *same scope*	1.9 (2/103)	7.2 (6/83)	
Reordered scope roles *out of scope*	12.62 (13/103)		

Table 7: Results from the error analysis of the 250 sentences randomly extracted from the IWSLT2014 test set.

A.3.6 Results: English-to-Korean

The result of the error analysis for the English-to-Korean translation task are shown in 8. At first glance we notice a substantial difference between these and the results for the Chinese-to-English tasks where the three categories shows similar scores. In particular cue translation shows a big drop in translation performance. The main reason for this is that English and Korean substantially differ in the way negation is expressed, which also justifies the presence of *partial* and *wrong* cues. The majority of translation errors classified as *partial cue* are cases where the cue was inflected in the wrong way (as in Ex. 10) but there are also other cases where the system failed to reproduce double negation on the target side (as in Ex. 11).

(10) a. *Src*: Which is^{+pres} **not**$_{\text{cue}}$ what you originally intend
b. *Hyp:* geos-eun wonlae gaecheogha-neunde **anieossseubnida**$_{\text{cue}}$.
Thing.TOP once to pioneer.CON not-to-be.**PAST**.HON

(11) a. *Src*: **No one**$_{\text{cue}}$ was allowed to speak about movement in plants before Charles Darwin
b. *Hyp:* Chalseu dawin-e jeon-e **amudo**$_{\text{cue}}$ umjig-im-eul sigmul-e daehae
Charles Darwin.LOC before.TMP no one move.NZR.ACC plants.LOC regarding speak.POT.HON
malhal su iss-eossseubnida$_{\text{event}}$.

[6] Given the poor quality of the reference translations, we could not carry out an analysis of the kind kind of negation instances and the tendencies in their translation for the present test set.

English-to-Korean: KAIST + TED - 250 sentences			
Average Sentence Length	12		
Number of negated sentences	31	12.4%	
Cue per sentence ratio		1.06	
	Src	Hyp	
Cues	33	31	
Events	33	22	
Scope roles	50	34	
	R%	*P%*	F_1
Correct cues	42.42 (14/33)	45.16 (14/31)	
+ Partial cues	51.51 (14 + **6**/33)	54.83 (14 + **6**/31)	**53.11**
Correct events	36.36 (12/33)	54.54 (12/22)	
+ Partial events	39 (12+ **2**/33)	59 (12 + **2**/22)	**46.95**
Correct scope roles	38 (19/50)	55 (19/34)	
+ Partial scope roles	41 (19 + **3**/50)	60 (19 + **3** /34)	**48.71**
Deleted cues	12 (4/33)		
Deleted events	12 (4/33)		
Deleted scope roles	4 (2/50)		
Inserted cue		9 (3/31)	
Inserted events		13.6 (3/22)	
Inserted scope roles		17.6 (6/34)	
Reordered events *out of scope*	30 (10/33)		
Reordered scope roles *out of scope*	40 (20/50)		
Reordered scope roles *same scope*	2 (1/50)	2 (1/50)	

Table 8: Results from the error analysis of the 250 sentences randomly extracted from the English-to-Korean test set.

Example (10) shows two main features of Korean negation: (i.) in a nominal predicate, negation is expressed by means of a negative copula, unlike English where the copula and negation cue are separate words; (ii.) verbal negation cues can be inflected. Here, the negative cue takes the wrong tense and it is therefore marked as partially correct. The hypothesis contains in fact **anieossseubnida** (+past) instead of **anibnida** (+pres). On the other hand, in (11), even though the negative pronoun is correctly translated, the system fails to negate the verb as well (it should be *malhal su **eobs-**eossseubnida* instead of *malhal su **iss-**eossseubnida*) . Problems in rendering double negation, where not present in the source, as also observed in Popovic and Arcan (2015), are expected given that the system is completely negation-agnostic and this phenomena cannot be captured by translation rules only.

Furthermore, unlike the Chinese-to-English pair, where cue - related errors were mostly related to deletion, in the Korean translations there were 8 instances of wrongly translated cues. Again, this is due to the fact that, while in English and Chinese there exists one predominant cue to express standard verbal negation (*not* and *bù* respectively), Korean has many, where the choice depends on associated eventive (nominal, existential, etc.) and aspectual features (potential, permission, experiential, etc.). These kind of errors are exemplified in (12) and (13).

(12) a. *Src*: [...] “OK, I did**n’t**$_{\text{cue}}$ really see that"
b. *Hyp:* [...] “geulae , naega boji **mos**$_{\text{cue}}$haessseubnida" .
[...] “Sure , I.SUBJ see.FIN-PART can-not.PAST" .

(13) a. *Src*: She reminded me that I had**n’t**$_{\text{cue}}$ written to Mother.
b. *Hyp:* geunyeoneun je eomeoniga jega sseun jeog-i **eobs-eoss-geodeun-yo**$_{\text{cue}}$
That woman.TOP my mother.SUBJ I.SUBJ to write.REL.PAST CL.SUBJ not-to-have.PAST.FIN-PART.POL

In (12), a wrong cue was chosen where Korean distinguishes between negation in potential form (equivalent to the English *can not*) where the cue *mos* is used, and plain negation (as in the English *do not*) where the cue *an* is used. In (13) instead a form expressing the lack of a certain experience (*-eun jeog-i eobs-eoss-geodeun-yo*, same as the English ‘I have never V+ed’) is incorrectly used in lieu of a standard negative form.

Beside some instances of cue deletion, that, as for the English-Chinese pair, are associated with wrong rule application and not with the presence of OOV items, 3 instances of cue insertion were found inspecting the Korean translations, where the source was a positive sentence. Using the trace of rule used during decoding, we verified that those are all due to the incorrect rule extraction during training, where a positive source phrase (or hierarchy of phrases) is associated with a negative one. This is shown in (14),

(14) a. *Src*: Squad cars converged on the scene of the crime.
b. *Hyp:* salam-ege chaleul converged beomjoehyeonjang-eun geuga hanbeondo **mos**gabon gos-ijiman
People.DAT car.ACC converged scene of the crime.TOP that.SUBJ not even once can **not** go.REL.PAST place to be.CON
geuui insaeng-eul yeong-wonhi bakkwonoh-eun .
that.POSS existence.ACC forever change.RES.PAST

where the source side *the scene of the crime* was directly into the target phrase *beom joehyeon jang-eun geuga hanbeondo **mos**gabon gos-ijiman*. Insertion errors can be attributed in the case to both data sparsity and the fact that we have not performed

any lemmatisation or morphological decomposition of the Korean side prior to alignment.

Event and scope related errors have similar patterns to the Chinese-to-English examples seen above. We have found that the biggest source of errors is movement outside of negation scope during translation. Out-of-scope reordering errors for both elements outnumber those errors where the element is kept inside the right scope but it is simply translated incorrectly (10 vs. 5 for the event and 20 vs. 6 for the scope). It is also worth reminding that we are considering here as *out of scope* also those elements that are translated into the target language when the cue is deleted (and a negation scope is therefore absent).

Finally, we found a few cases of event and scope role insertion associated to the appearance of a negative instance on the target side where the source side does not contain any. This is the case of (15), where the mistranslation of a positive sentence into a negative one leads to the appearance of an event and scope roles.

(15) a. *Src*: I can say with certainty that this is true.

b. *Hyp:* [jeoneun]$_{\text{scope.role}}$ [igeos-i]$_{\text{scope.role}}$ [sasil-ilaneun geos-eul]$_{\text{scope.role}}$ [hamkke]$_{\text{scope.role}}$ [malhal su **eobsda**$_{\text{cue}}$]$_{\text{event}}$.
I.TOP this thing.SUBJ truth.to-be.PART.REL.PRES thing.ACC together to speak.REL CL not-have.PL

When investigating the kind of negation instances for the English-to-Korean sentence pairs, out of the 33 negation instances found in the test set, we found that only three are of a negation outside a VP. This is very similar to what we observed in the case of the Chinese-to-English translation task. Analysing whether both source and reference translation realise the negated VP in the same way, we found that only in 2 cases a verbal VP is rendered by a nominal VP and 2 other cases where a verbal VP is translated using a positive paraphrase, leaving therefore 26 cases of 1-to-1 correspondence between the type of VP source and reference translation realise negation in.

Finally, we also found a strong 1-to-1 correspondence when analysing the sentential environment negation is realised in both source and reference translations. We found in fact only two cases of source-side negation in a subordinate that it is promoted to the main clause in the reference translation.

A.3.7 Discussion

Given the results for both translation tasks, there are a few point worth highlighting. First, the choice of language pair influence translation performance and the errors observed. Translating into languages that express negation morphologically or apply morphological operations to the cue adds more variables to the problem of translating negation. Second, for distant language pairs (such as the ones we have analysed here) there is a problem with the reordering of elements. This effects the scope of negation in particular whose semantic sub-constituents are translated outside of it.

Finally and more in general, we have seen how applying a small set of string based operations on a finite class of semantic elements allowed us to analyse the translation of textual negation throughly.

To summarise the results, for each of elements considered we might want to take into consideration that:

- cue: deletion and choice are the most prominent problems for this element. Deletion is not desirable because if the cue is absent, the truth value of the statement is reversed, hence we need a way to ensure that is present. Choice both refers to the type of cue and the morphological features it has to be specified for. This was evident especially in the Korean translations, where the choice of a correct cue and its inflection is not a trivial task.
- event: we have observed all categories of errors for the event, including deletion, out-of-scope movement and incorrect translation. We have also seen that deletion is not caused by OOV items in the majority of cases, which then leads to the conclusion that it if the event is absent is mainly because translation rules haven't promoted it correctly.
- cue & event: given that most of the negation instances observed are situated inside a VP and likely to be translated inside a VP in the target language as well, the issues related to the translation of cue and event can be brought together. Given the error observed, our goal is to ensure that we reproduce a predicate containing cue (or better, the correct cue) which is positioned correctly in a constituent dominated by the right event head. This is not a trivial task in both target languages we have considered: English requires that negation is inserted correctly inside a chain of auxiliaries or modals that in turn take on certain morphological features, whereas Korean requires that the cue is correctly placed in the chain of morphemes attached to the main verb (when morphological).
- scope: the main problem we have observed is out of scope reordering caused by rule application that are agnostic to clause boundary and predicate frames.

A.3.8 Chapter summary and future work

In the present section, we have presented a manual error analysis of the translation output with regards to negation. String-based operations (deletion, insertion and reordering) were used to score the translations of the three elements of negation introduced in §2 via a precision-recall based metric. Output from a Chinese-to-English and an English-to-Korean shown that language similarity in expressing negation matters, especially in relation to the translation of the cue. Scope re-ordering was found to be a consistent problem in both systems; this might be due to unconstrained rule application in HIERO systems. Finally, we have found that for the test sets used, the majority of negation instances are placed around a predicate VP.

A question related to the error analysis is whether it is possible to automate it. Given that we cannot rely on human judgement, an automatic error analysis will have to take into consideration the reference translation instead of the source. Understanding how much the machine output differs from the gold standard reference overlaps with the question of how to automatically evaluation the translation output.

A.4 Automatic evaluation of negation related errors in translation

Given the manual error analysis sketched in §3, we propose the following plan for an automatic semantic evaluation of negation related errors in translation:

- Automatically detect cue, event and scope in the reference translation
- Try to adapt this automatic detection model to other languages
- Automatically score the overlap between cue, event and scope in the reference and the machine output

The following section will cover 1) and partially 2). We leave 3) for future work.

A.5 Automatic negation detection

If the task of the present project is to guide the system to avoid as much as possible the errors mentioned in §2, we have first of all to be able to recognise on the source side the elements of negation to which the errors are related to.

Detecting these elements manually is time-consuming, especially if experiments on different test sets and language pairs are to be carried out. Therefore, it would be useful to have a method to automatically detect these elements that can scale to different domains and different languages.

In this section, the possibility of leveraging available training data annotated for negation in English is investigated. The problem of automatically detecting negation is framed as a supervised machine learning task, the output of which can then be projected onto a target foreign language in the presence of aligned parallel data. Given that previous literature has already explored the problem of detecting negation in English texts (§4.1), we will try to re-implement a pre-existing algorithm for negation detection (§4.2) and port them to the newswire domain (§4.3) where parallel data is available to perform annotation projection.

A.5.1 Previous work

Previous literature on automatic negation detection has focused mainly on its use in sentiment analysis and in the biomedical domain. To our knowledge, all these works have tackled the problem using either a rule-based approach based on NegEx (Chapman *et al.*, 2001) and its extensions or supervised machine learning methods.

NegEx is a simple and fast rule-based algorithm that is able to detect the scope of a negation cue in the context of medical referrals. Here, scope is intended as one or more NE that are often associated with a disease or symptom, such as *chest pain* in the sentence 'The patient denied experiencing chest pain on exertion'. The task NegEx was initially applied to was to recognise whether a NE is negated or not. All these NEs are part of a precompiled list of disease and symptoms.

Since its basic implementation, NegEx has been extended and adapted to other languages. (Harkema *et al.*, 2009) extends the scope of NegEx to match information about the temporal context of the disease or symptom, whether these are factual or not and whether they concern the patient or a third person. (Chapman *et al.*, 2013) tries to extend the NegEx lexicon to languages other than English by translating the NegEx cue dictionary in three European languages (French, German and Swedish) and create knowledge representations based on multi-lingual ontologies. Despite its extensions, NegEx has however been restricted only to the domain of medical records, with specifically tailored rules; moreover the process of porting the algorithm to other languages has been carried out by means of manual translations, whereas we want to explore the option of automatising the process.

Until very recent all supervised learning approaches were relying on the BioScope corpus (Szarvas *et al.*, 2008) and other few resources in the biomedical domain. The BioScope corpus is a collection of clinical texts (e.g. radiology reports), biological reports and scientific abstracts annotated for both negation and hedging. The annotation only takes into consideration the cue with its related scope; there is no separate annotation of the event, which is considered as part of the scope. An excerpt from the BioScope corpus is shown in (16).

(16) <sentence id="S1.15">Our result <xcope id="X1.15.3"><cue type= "speculation" ref="X1.15.3">suggests </cue>that <xcope id= "X1.15.2" >the unknown amino acid encoded by stop codons does <xcope id= "X1.15.1"><cue type ="negation" ref ="X1.15.1 " >**not** </cue>**exist** </xcope>, <cue type ="speculation" ref ="X1.15.2" >or </cue >its phylogenetic distribution is rather limited </xcope ></xcope >, which is in agreement with the previous study on tRNA. </sentence>

It can be observed in (16) that what we consider as scope in the error analysis (underlined above) is not considered as such in the BioScope corpus; this applies to all elements outside the negated VP, which are not considered in the scope of negation (although, erroneously, 'does' is excluded from the scope as well). Moreover the cue is considered as part of the scope as opposed to separate from it.

In sentiment analysis, (Councill *et al.*, 2010) has successfully used the BioScope corpus to both train and test a negation detection algorithm that is then inserted as part of a sentiment recognition system, showing considerable improvement in discriminating between positive and negative polarity. On the other hand, other works (Ballesteros *et al.*, 2012; Zou *et al.*, 2013) focus purely on detecting the scope of negation using deep syntactic features. Finally, (Prabhakaran and Boguraev, 2015) highlight the inconsistencies in the BioScope annotations and proposes a mapping from annotations to predicate-argument structures before carrying out the prediction task.

If the NegEx algorithm and the BioScope corpus were developed with the aim of automatically extracting negation related information in the biomedical domain, the *SEM 2012 shared task represented a first attempt to detect negation outside this domain. In particular, extracts from Conan Doyle's "Sherlock Holmes" were annotated in CoNLL format for cue, scope and event according to the guidelines introduced in §2. An excerpt from the annotation is presented in (17).

(17)
```
baskervilles01 8 0 Holmes Holmes NNP (S(S(NP*)                    _          _          _
baskervilles01 8 1 was be VBD (VP*                          _          _          _
baskervilles01 8 2 sitting sit VBG (VP*                       _          _          _
baskervilles01 8 3 with with IN (PP*                        _          _          _
baskervilles01 8 4 his his PRP$ (NP*                        _          _          _
baskervilles01 8 5 back back NN *))                         _          _          _
baskervilles01 8 6 to to TO (PP*                      _          _          _
baskervilles01 8 7 me me PRP (NP*)))))                      _          _          _
baskervilles01 8 8 , , , *                      _          _          _
baskervilles01 8 9 and and CC *                       _          _          _
baskervilles01 8 10 I I PRP (S(NP*)                   _          I          _
baskervilles01 8 11 had have VBD (VP*                       _          had        _
baskervilles01 8 12 given give VBN (VP*                       _          given      given
baskervilles01 8 13 him him PRP (NP*)                       _          him        _
baskervilles01 8 14 no no DT (NP(NP*                  no         _          _
baskervilles01 8 15 sign sign NN *)                   _          sign       _
baskervilles01 8 16 of of IN (PP*                     _          of         _
baskervilles01 8 17 my my PRP$ (NP*                   _          my         _
baskervilles01 8 18 occupation occupation NN *))))))        _          occupation
baskervilles01 8 19 . . . *)                    _          _          _
```

All the sentences in the story are annotated with story id, sentence id, word id, surface form, lemma, POS tag and constituent fragment. Additionally, if the sentence is negated each negation instance is represented by three column containing cue, scope and event respectively. The fact that the event (*given* in (17)) is included in the scope explains why a word-event is also present in the second column.

The evaluation of the automatic recognition task is precision and recall based. The task is to correctly identify the negated sentences and correctly detect cue, event and scope. Given the goal of the task, all of the system submitted for the scope resolution task[7] attempted first to detect the cue and given this, to then detect scope and event. In particular CRFs and SVMs, making use of syntactic (both constituent and dependency based) features, were shown to lead to the best results in a supervised machine learning setting. Three of the systems submitted, one of which we re-implement in the next section, show different strategies on how the scope can be predicted: (Read *et al.*, 2012) observed that the scope often matches a syntactic constituents and it is therefore a matter of choosing the right constituent that can capture the scope of a particular sentence. For this reason, they use discriminative ranking on candidate constituents to choose the most appropriate to capture the scope of a negation instance; on the other hand, (Lapponi *et al.*, 2012) used features extracted from a dependency parse to detect elements in the scope, such as the dependency path from a given word to the cue and the parent and the children tokens and dependency labels (see next section for more details); (Basile *et al.*, 2012) tried instead to detect the the scope in a DRS representation (Discourse Representation Structure), including all arguments around a negated item. It is worth pointing out that no matter it is constituents, dependency structures or semantic representations we are taking into consideration, there is an common element of 'cohesiveness' where the elements in the scope are likely to be associated with the boundaries of the sentential environment the cue is positioned in.

Beside English, there were also a few attempts in automatically detecting negation in Chinese texts. (Hao-Min *et al.*, 2008) designed a negation detection algorithm based on syntactic patterns; similarly, (Jia *et al.*, 2014) implemented an FSA for automatic recognition of negation structures in Chinese medical texts, using a list of manually defined cues and the syntactic structures they appear in.

The problem of annotating languages where no manually annotated training data is available resembles very closely the task of building SRL systems in low-resources languages. We are in fact faced here with the task of having a strong supervised model on a language (that we denote here as the *source* language) and wanting to exploit parallel data to transfer this knowledge to a foreign language (the *target* language). Previous work on SRL have proposed three different approaches to the problem: (i.) annotation projection; (ii.) direct model transfer and (iii.) unsupervised learning (which we will leave aside for the moment).

The intuition behind the *annotation projection* approaches is that, in the presence of either manually or automatic aligned parallel data, we can project the annotations from the source onto the target language. (Padó and Lapata, 2005) were the first to use word-alignment information for cross-lingual projection of FrameNet annotations from English onto German. However, given that word alignments are noisy and that certain words might not be aligned, generating discontinuities that are not allowed in SRL, projection is enhanced using syntactic constituent information as well. To this regard, (Padó and Lapata, 2006) casts the problem of using syntactic constituents in SRL annotation projection as weighted graph optimization, where we want to find the best alignment between source and target constituents. Although experiments were limited to a single language pair and to those sentences where frames match cross-linguistically, the possibility of transferring roles by means of a simple heuristic is interesting. On the other hand, (Van der Plas *et al.*, 2011) attempts to remedy to incorrect word alignments by first using word-based projection to transfer semantic information to the target language (in this case French) and then use this, along with a dependency parse on the target side, to train a joint syntactic-semantic parser.

In the *direct model transfer* (Kozhevnikov and Titov, 2013) approach, a model is built on the annotated source language and directly applied to the target language. Direct application implies that the model features are abstract enough to be applicable to both languages. (Kozhevnikov and Titov, 2013) experiments with different feature representations such as universal POS tags, unlabeled dependencies and cross-lingual word mapping. In this last case, we are able to preserve token-related features by either mapping target tokens to source tokens (*glossing*) or mapping both to a cross-lingual word cluster. Results show performance comparable to the annotation projection approach but with no need for parallel data.

[7] The shared task also included a focus resolution task that we are not considering since outside the scope of the present work

List of English lexical negation cues	List of English morphological cues
nor	im-
neither	un-
without	dis-
nobody	in-
none	non-
nothing	ir-
never	-less
not (n't)	
no	
nowhere	
non	

Table 9: List of English lexical and morphological negation cues

A.5.2 A pipeline for automatic negation detection

Given that algorithms for automatically detect negation already exist, the most sensible starting point would be to re-implement one of them and test its performance on the task at hand, i.e. detect the elements of negation in the newswire domain.

The reason why we chose to re-implement (Lapponi *et al.*, 2012)'s system is mainly because dependency parse features have often proved useful when dealing with negation (see for example (Prabhakaran and Boguraev, 2015) and (Fancellu and Webber, 2014)); moreover, unlike, constituent based algorithms (Read *et al.* (2012)), working with dependency parses requires less engineering while still being applicable to a wide range of languages.

As introduced in the section above, (Lapponi *et al.*, 2012)'s system is based on two sequential classifier, one for cue and the other for scope detection.

Cue detection is a binary classification task: negation cues are a closed class and the task is to identify whether a target word belong to this class or not. This approach applies to both lexical and morphological negation in English. As for lexical negation, we start from a finite set of elements (shown on the left in 9) and try to classify whether a word is a true cue or an instance of non-functional negation (e.g. *not* in question tags). In the case of morphological negation, we look for words containing a sub-string that can potentially be an affix expressing negation (such as **im-** in *impossible*, see liston the left in 9) and our task is *not* to classify as negation-bearing those words containing a substring that is the same as the negation cue but has no negation related meaning at all (as the *im* in *impostor*). To carry out this task, an SVM classifier is trained using the set of features reported in (18).

(18)
- Lexical cues:
  - Lemma n-grams to the left of cue (up to 5-gram)
  - Lemma n-grams to the right of cue (up to 5-gram)
  - Token n-grams to the left of the cue (up to 5-gram)
  - Token n-grams to the right of the cue (up to 5-gram)
- Morphological cues:
  - Lemma n-grams to the left of cue (up to 5-gram)
  - Lemma n-grams to the right of cue (up to 5-gram)
  - Token n-grams to the left of the cue (up to 5-gram)
  - Token n-grams to the right of the cue (up to 5-gram)
  - Backward stem character n-grams
  - Forward stem character n-grams
  - Frequency of the stem in the training corpus (cues excluded).
  - Affix
  - POS tags

Some of the features in (18) tries to capture the context around the word; for instance, we will expect the present of a backward unigram *does* and a forward bigram (*it, ?*) to help classify those non-functional true negatives in question tags.

An additional set of features targets directly the disambiguation of words bearing morphological cues using the stem, i.e. the token after the potential negation morpheme has been stripped (e.g. *possible* from *im*possible). If the morpheme is a real negation cue that derives a word bearing negative meaning from a positive one, it is likely that the stem appears on its own in the training corpus. Hence, we take away all the cues from the training data and calculate the frequency of the stem in it. Following this intuition, if we strip **im-**, a real negation cue, from *impossible*, the stem *possible* is very likely to appear on its own in the 'negation-less' copy of the training corpus. On the other hand, in the case of a word like *underground*, which contains a false negation morpheme *un-*, its stem *derground* is not a real word and therefore very unlikely to appear in the training corpus. For the same reason, stem character n-grams, both beginning-to-end (forward) and end-to-beginning (backwards) are considered as well. Finally, the cue detection algorithm also includes a post-processing component responsible for detecting multi-word

negation cues that do not need disambiguation and can therefore be detected by a simple string-matching lookup.

The output of cue prediction gives us a way to detect which are the negated sentences and how many instances of negation appear in a sentence. Given the cue, we can then detect the scope and the event in it. Unlike the cue detection task where we had to discriminate real vs. false cues, the scope (and jointly event) detection task is an IOB (Inside-Outside-Beginning) classification problem where we want to classify a target token as whether part of a negation scope or not. In order to carry out the task, a CRF (Conditional Random Field) is trained using the set of features reported in 10. If the set of general features

General features
Token
Lemma
PoS unigram
Forward token bigram and trigram
Backward token bigram and trigram
Forward PoS trigram
Backward PoS trigram
Lexicalized PoS
Forward Lexicalized PoS bigram
Backward Lexicalized PoS bigram
Constituent
Dependency relation
First order head PoS
Second order head PoS
Lexicalized dependency relation
PoS-disambiguated dependency relation
Cue-dependent features
Token distance
Directed dependency distance
Bidirectional dependency distance
Dependency path
Lexicalized dependency path

Table 10: List of the features using to train the CRF model for scope detection as appearing in (Lapponi *et al.*, 2012, p. 322)

encode the environment around a given token, the cue-dependent features try to use the cue as an ulterior guidance towards tracing the boundary of the scope. Both dependency distance features represent how far in the dependency graph is the target element from the cue; on a string level, this is calculated by the token distance.

(Lapponi *et al.*, 2012)'s model is implemented using the Wapiti toolkit (Lavergne *et al.*, 2010), with default settings. Here, we use the Mallet toolkit (McCallum, 2002) with default settings. The output classes considered from this task are O (*outside* the scope), S (inside the *scope*), B (*beginning* of the scope), E (event) along with C (lexical cue) and MC (morphological cue), predicted in the previous task. (Lapponi *et al.*, 2012) opted to separate lexical and morphological cues given the different context they appear in.

The scope classification task also includes a post-processing step that deals with the fact that a sentence might contains multiple negation instance, sometimes nested (as in the case of a negated subordinate inside a negated main clause, as in [She *is* **not** *coming* because [her dad *is* **not** *feeling* well]]), while the CRF classifier just assigns a inside vs. outside value to a test token. In order to solve potential nesting (in this case whether cue B is nested in the scope of cue A), (Lapponi *et al.*, 2012)(p. 324) uses the following heuristics:

- Cue B is to the right of A.
- There are no tokens labeled with S between A and B.
- Token distance between A and B does not exceed 10.

Once the cue hierarchy has been decided, each token that has been predicted as been part of a scope is assigned to a cue according to the heuristics below.

- Assign each token T to the closest negation cue A with no S-labeled tokens or punctuation separating it from T.
- If A was found to be negated by cue B, assign T to B as well.
- If T is labeled with E by the event classifier, mark it as an event.

The result for the cue classification tasks are reported in Table 11. First, it is interesting to observe how disambiguating between cues achieves a high F_1, with detection of lexical cues performing slightly higher than the disambiguation of morphological negation. If we combine the prediction of lexical and morphological cues and add post-processing rules to automatically detect multi-word cues, we achieve a final F_1 score of 0.92. An analysis of the confusion matrices for both test sets reveal that in

		P	R	F_1
Cardboard	Lexical	0.95	0.94	0.93
	Morphological	0.92	0.93	0.92
Circle	Lexical	0.94	0.94	0.93
	Morphological	0.91	0.91	0.91
Average	Lexical	0.945	0.95	**0.93**
	Morphological	0.915	0.92	**0.915**

Table 11: Results for cue classification task on both tests sets released during the *SEM2012 task.

		P	R	F_1	B	E	O	S
Cardboard	Beginning	0.79	0.73	0.76	79	1	24	4
	Event	0.48	0.62	0.54	0	45	8	19
	Outside	0.98	0.99	0.99	10	2	9089	66
	Scope	0.85	0.74	0.79	11	46	122	515
Circle	Beginning	0.73	0.69	0.71	66	1	24	5
	Event	0.49	0.59	0.54	1	46	9	22
	Outside	0.97	0.99	0.98	13	4	7978	75
	Scope	0.80	0.66	0.72	10	41	164	422
Average	Beginning	0.76	0.71	**0.735**				
	Event	0.485	0.605	**0.54**				
	Outside	0.92	0.93	**0.985**				
	Scope	0.825	0.70	**0.755**				

Table 12: Results for scope classification task, along with the confusion matrices, of both test sets released during the *SEM2012 task

the case of lexical negation for both test sets, classification errors involves only false positives (7 out of 12 true negatives), that is those non-functional cues that do not carry a negative meaning. In the case of morphological negation, we observe instead both false positive and false negative. Part of the errors associated with this class are related to words that do not contain a negative morpheme but exist as a relatively high-frequency word if stripped of the substring resembling a negation morpheme (e.g. 'in*land*'); on the other hand, other words that contained a negation morpheme were not classified as true positives because the stems is a low-frequency word and therefore not present in the training data as separate word (e.g. '**un**framed').

The result for the scope classification task are reported in Table 12. In terms of F_1 we notice a good performance for the scope token recognition but not for the event recognition (when the event is not part of a word with a negation morpheme). It is worth noting however that for the task of scope classification an investigation of the confusion matrix is important, since output from the B class are in reality a subcategory of S and S outputs that are classified as E are still to be considered correct (because, as we said, the event is included in the scope). For this reason we re-ran the classification task without the B and the E class. Results are reported in Table 13. It can be observed that by making the task a binary classification (whether the token is inside or outside the scope) the F_1 measure for the S category reaches a solid 0.85 on average. As for the event, which is part of the scope, we hypothesize that its detection can be performed taking the output of the scope detection as input, as so to assure that a token outside the scope is not picked. This will be addressed by future work.

The results shown so far only report the accuracy in classifying whether a token is inside or outside the scope. As shown above, if these results are positive for those sentences containing only a single instance of negation or multiple instances that are not nested, in the case of sentences with nested negation, there exists the extra task of assigning the scope to the right cue. To report the results of this task using the heuristics of (Lapponi *et al.*, 2012) introduced above, we use the evaluation script released during the *SEM2012 shared task for the scope evaluation task. The script evaluates whether a full scope is captured correctly,

		P	R	F_1	O	S
Cardboard	Outside	0.99	0.99	0.99	9065	101
	Scope	0.88	0.88	0.88	106	768
Circle	Outside	0.98	0.99	0.98	7974	95
	Scope	0.86	0.77	0.82	183	627
Average	Outside	0.985	0.99	**0.985**		
	Scope	0.87	0.825	**0.85**		

Table 13: Results for scope classification task, along with the confusion matrices, of both test sets released during the *SEM2012 task when only two output classes are considered

		P	R	F_1	O	S
Cardboard	Outside	0.98	0.98	0.98	9076	90
	Scope	0.89	0.87	0.88	116	758
Circle	Outside	0.98	0.98	0.98	7928	142
	Scope	0.81	0.79	0.80	172	639
Average	Outside	0.98	0.98	**0.98**		
	Scope	0.85	0.83	**0.84**		

Table 14: Results for scope classification task, along with the confusion matrices, of both test sets released during the *SEM2012 task when only token and POS features are retained.

		P	R	F_1	O	S
Cardboard	Outside	0.98	0.98	0.98	9023	144
	Scope	0.83	0.83	0.83	151	723
Circle	Outside	0.97	0.97	0.97	7863	207
	Scope	0.73	0.73	0.73	222	589
Average	Outside	0.975	0.975	**0.975**		
	Scope	0.78	0.78	**0.78**		

Table 15: Results for scope classification task, along with the confusion matrices, of both test sets released during the *SEM2012 task when only POS features are retained.

also taking into consideration whether we match it to the correct cue or not. The F_1 measure for the full scope condition is 65.67 with a precision and a recall of 90.41 and 51.56 respectively. This is around 7 points lower than the original implementation; there might be part of the scope assignment implementation that needs revisiting, which we will address in future work.

So far, we have showed that recognising the cue and the elements of the scope in sequential fashion leads to a relatively high F_1 measure on in-domain data. Before assessing whether such results also hold for out-of-domain data, we show the results of an ulterior experiment we carried out. The question we wanted to address is whether a simpler model that would show a similar performance could be created; while addressing this question, we also tried to explore the possibility of moving towards language-independent features, that would make the model directly applicable to other languages. To this purpose, we started by performing *feature ablation* and considering only token and POS related features. Results are shown in Table 14. Although there is a slight drop in performance, the F_1 score for scope detection is still above 0.80, showing how token and POS related information are essential in defining the boundaries of a negation scope. We then proceeded in considering POS-related features only; considering that universal POS sets are available for a large number of languages, if the performance of the classification task is still good, it might be worth considering to train a POS-based model for English and directly apply it to other languages. The results for this experiment are shown in Table 15. When analysing the results for the classification task using POS-based features only, performance is still around 0.80 as measured by the F_1 score, showing that the core of the classification task is done in terms of POS tags.

A.5.3 Automatic negation detection on newswire data

Going back to the original problem, we are trying to build a detection algorithm that is able to scale across domains and language pairs. Given that machine translation often makes use of newswire and web data, it is worth investigating how a negation detection algorithm performs on these new domains. Moreover, given the availability of parallel data, some of which are manually aligned, it is possible to go beyond automatic annotation and project negation related information from a language to another. To this purpose, the negation detection pipeline shown in the previous section is applied to the English-to-Chinese GALE manually aligned parallel data (LDC2012T24). The advantage of using a manually aligned data is two-fold: cross-lingual projections do not suffer from any noise derived from automatic alignment; manual alignment can be used as a benchmark to assess the performance of the projection task using automatic alignment heuristics.

As a first experiment we train our automatic negation detection system on the Conan Doyle corpus and test it on the English side of the GALE Chinese-English manually aligned parallel corpora. Before the annotation process, the 4842 sentences had tokenised and formatted into ConLL format; tokens were lemmatised, part of speech tags and constituent fragments extracted using the Stanford CoreNLP toolkit (Manning *et al.*, 2014). We detect cue and scope using the classifiers shown in the previous section with the entire set of features (token, POS and dependency based); a gold standard for evaluation was finally created by correcting the classification output. We carried out this correction on the first 1000 sentences in the corpus, following the same guidelines used during the manual error analysis. The results in Table 16 refers to this subset of 1000 sentences.

A.5.4 Chapter summary and future directions

In the present chapter, we have explored the issue of automatically detecting the sub-constituents of negation in the source sentence in order to discern which elements we have to guide the translation of. We have shown that by re-implementing

	P	R	F_1	gold	system	tp	fp	fn
Cue	86.38	92.69	**89.42**	301	333	279	44	22
Scope tokens	77.17	72.46	**74.74**	2556	2400	1852	548	704
Full scope (cue match)	74.23	40.07	**52.05**	302	326	121	42	181
Full scope (no cue match)	75.29	42.38	**54.23**	302	326	121	42	174

Table 16: Result for automatic detection of negation on the English side of the GALE English-Chinese parallel corpora.

algorithms previously developed for English it is possible to obtain good performance on cue and scope detection. We have also looked at the contribution of different features in the scope classification task in order to assess the contribution of each. By means of *feature ablation*, we have observed that the main contribution comes from POS-tag related features, while adding dependency features on top helps improving the overall performance. Finally, the issue of testing on a genre different then the training data was explored; performance worsens slightly when considering newswire data on both cue and scope detection.

Future work will focus on: (i.) further analysing the results and improving on what done so far; (ii.) tackling the problem of automatically detecting negation in other languages. In the case of (i.), future tasks include:

- *Predicting the event*. We have observed that predicting the event alongside scope elements leads to some instances being predicted as out of scope. Given that the event is always inside the scope, we will explore the possibility of first predicting the scope and then using this prediction as input to an event classifier.
- *Improving scope detection using semantic features*. In the experiments here reported, we have only used word-based and syntactic features to carry out the scope prediction task. Recent work by (Packard *et al.*, 2014) shows that it is possible to outperform all systems submitted for the *SEM2012 scope detection task by using a rule-based heuristics on an MRS (Minimal Recursion Semantics) graph; we will then try to re-implement such heuristics and apply it to the GALE data in order to investigate whether performance compares also across genres.
- *Testing with different feature combination on the GALE corpus*. Negation detection on the English side of the GALE corpus has been carried out using the full set of features, under the assumption that this combination lead to the best performance on the Conan Doyle test data. However, for completeness, it is important to perform feature ablation also on the GALE corpus to make sure this applies also on out-of-genre documents.
- *Analysing the errors in the GALE corpus*. We have seen that in terms of F_1 scores, performance worsens slightly when testing on the GALE English data. We haven't however analysed what are the reasons of this drop in the performance of both cue and scope classification.

As for (ii.), we will proceed by:

- *Building a model with language-independent features*. We have seen that by using only token and POS-related features, the performance of the system does not worsen, which is relevant if we are to train a model on the source English side and directly transfer it onto a foreign language, following the work of (Kozhevnikov and Titov, 2013). We will experiment with the same set of universal features: cross-lingual word clusters, unlabeled dependencies and universal POS tags.
- *Performing annotation projection through word alignments*. Given the availability of a manually aligned parallel corpus, it is possible to build a baseline by projecting the annotations from English to Chinese using word alignment information only and using the alignment of syntactic constituents as a way to improve from this baseline. This is motivated by work previously done in semantic role labelling but also by the fact that syntactic constituents are good predictors of scope boundaries (Read *et al.*, 2012).

References

Baker, Kathryn, Michael Bloodgood, Bonnie J Dorr, Chris Callison-Burch, Nathaniel W Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. "Modality and negation in simt use of modality and negation in semantically-informed syntactic mt." *Computational Linguistics*, 38(2):411–438.

Ballesteros, Miguel, Virginia Francisco, Alberto Díaz, Jesús Herrera, and Pablo Gervás. 2012. "Inferring the scope of negation in biomedical documents." *Computational Linguistics and Intelligent Text Processing*, 363–375. Springer.

Basile, Valerio, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. "Ugroningen: Negation detection with discourse representation structures." *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 301–309.

Bazrafshan, Marzieh and Daniel Gildea. 2013. "Semantic roles for string to tree machine translation." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 419–423. Sofia, Bulgaria.

Birch, Alexandra and Miles Osborne. 2011. "Reordering metrics for mt." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1027–1035. Portland, Oregon, USA.

Blanco, Eduardo and Dan I Moldovan. 2011. "Some issues on detecting negation from text." *FLAIRS Conference*.

Bojar, Ondrej, Rudolf Rosa, and Aleš Tamchyna. 2013. "Chimera–three heads for english-to-czech translation." *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 90–96.

Bojar, Ondřej and Aleš Tamchyna. 2015. "CUNI in WMT15: Chimera Strikes Again." *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 79–83. Lisboa, Portugal.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. "Wit3: Web inventory of transcribed and translated talks." *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, 261–268.

Chapman, Wendy W, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. "A simple algorithm for identifying negated findings and diseases in discharge summaries." *Journal of biomedical informatics*, 34(5):301–310.

Chapman, Wendy W, Dieter Hilert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, Michael Conway, Melissa Tharp, Danielle L Mowery, and Louise Deleger. 2013. "Extending the negex lexicon for multiple languages." *Studies in health technology and informatics*, 192:677.

Chiang, David. 2007. "Hierarchical phrase-based translation." *Computational Linguistics*, 33(2).

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Collins, Michael, Philipp Koehn, and Ivona Kučerová. 2005. "Clause restructuring for statistical machine translation." *Proceedings of the 43rd annual meeting on association for computational linguistics*, 531–540.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. "Natural language processing (almost) from scratch." *J. Mach. Learn. Res.*, 12:2493–2537.

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. "Minimal recursion semantics: An introduction." *Research on Language and Computation*, 3(2-3):281–332.

Councill, Isaac G, Ryan McDonald, and Leonid Velikovich. 2010. "What's great and what's not: learning to classify the scope of negation for improved sentiment analysis." *Proceedings of the workshop on negation and speculation in natural language processing*, 51–59.

Fancellu, Federico. 2013. *Improving the performance of Chinese-to-English Hierarchical Phrase Based Models (HPBM) on Negative data using n-best list re-ranking*. Master's thesis, School of Informatics, University of Edinburgh.

Fancellu, Federico and Bonnie Webber. 2014. "Applying the semantics of negation to smt through n-best list re-ranking." *EACL 2014*, 598.

Fancellu, Federico and Bonnie Webber. 2015. "Translating negation: A manual error analysis." *ExProM 2015*, 1.

Galley, Michel, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. "What's in a translation rule?" *HLT-NAACL '04*.

Gao, Qin and Stephan Vogel. 2008. "Parallel implementations of word alignment tool." *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, 49–57.

Gong, Zhengxian, Min Zhang, Chewlim Tan, and Guodong Zhou. 2012. "N-gram-based tense models for statistical machine translation." *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 276–285.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. "Announcing Prague Czech-English Dependency Treebank 2.0." *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, 3153–3160. Istanbul, Turkey.

Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. 2006. "Prague Dependency Treebank 2.0." LDC2006T01, ISBN: 1-58563-370-4.

Hao-Min, Li, Ying Li, Hui-Long Duan, and Xu-Dong Lv. 2008. "Term extraction and negation detection method in chinese clinical document." *Chinese Journal of Biomedical Engineering*, 27(5).

Hardmeier, Christian and Marcello Federico. 2010. "Modelling pronominal anaphora in statistical machine translation." *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*, 283–289.

Harkema, Henk, John N Dowling, Tyler Thornblade, and Wendy W Chapman. 2009. "Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports." *Journal of biomedical informatics*, 42(5):839–851.

Jang, Chung-Hyok and Kwang-Hyok Kim. 2015. "The improvement of negative sentences translation in english-to-korean machine translation." *arXiv preprint arXiv:1512.08066*.

Jia, Zheng, Haomin Li, Meizhi Ju, Yinsheng Zhang, Zhenzhen Huang, Caixia Ge, and Huilong Duan. 2014. "A finite-state automata based negation detection algorithm for chinese clinical documents." *Progress in Informatics and Computing (PIC), 2014 International Conference on*, 128–132.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, *et al.* 2007. "Moses: Open source toolkit for statistical machine translation." *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180.

Kozhevnikov, Mikhail and Ivan Titov. 2013. "Cross-lingual transfer of semantic role labeling models." *ACL (1)*, 1190–1200.

Lapponi, Emanuele, Erik Velldal, Lilja Øvrelid, and Jonathon Read. 2012. "Uio 2: sequence-labeling negation using dependency features." *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 319–327.

Lavergne, Thomas, Olivier Cappé, and François Yvon. 2010. "Practical very large scale crfs." *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 504–513.

Li, Jin-Ji, Jungi Kim, Dong-Il Kim, and Jong-Hyeok Lee. 2009. "Chinese syntactic reordering for adequate generation of korean verbal phrases in chinese-to-korean smt." *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 190–196.

Li, Junhui, Philip Resnik, and Hal Daumé III. 2013. "Modeling syntactic and semantic structures in hierarchical phrase-based translation." *Proceedings of NAACL-HLT*, 540–549.

Lo, Chi-kiu and Dekai Wu. 2011. "Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 220–229.

Lopatková, Markéta, Zdeněk Žabokrtský, and Václava Benešová. 2006. *Valency Lexicon of Czech Verbs VALLEX 2.0*. Tech. Rep. 34, UFAL MFF UK.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. "The stanford corenlp natural language processing toolkit." *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

McCallum, Andrew K. 2002. "{MALLET: A Machine Learning for Language Toolkit}."

Mi, Haitao, Liang Huang, and Qun Liu. 2008. "Forest-based translation." *Proceedings of ACL-08: HLT*, 192–199. Columbus, Ohio.

Miller, George A. 1995. "Wordnet: a lexical database for english." *Communications of the ACM*, 38(11):39–41.

Morante, Roser and Eduardo Blanco. 2012. "* sem 2012 shared task: Resolving the scope and focus of negation." *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 265–274.

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

Och, Franz Josef. 2003. "Minimum error rate training in statistical machine translation." *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 160–167.

Packard, Woodley, Emily M Bender, Jonathon Read, Stephan Oepen, and Rebecca Dridan. 2014. "Simple negation scope resolution through deep parsing: A semantic solution to a semantic problem." *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics*.

Padó, Sebastian and Mirella Lapata. 2005. "Cross-linguistic projection of role-semantic information." *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 859–866.

Padó, Sebastian and Mirella Lapata. 2006. "Optimal constituent alignment with edge covers for semantic projection." *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 1161–1168.

Popovic, Maja and Mihael Arcan. 2015. "Identifying main obstacles for statistical machine translation of morphologically rich south slavic languages."

Prabhakaran, Vinodkumar and Branimir Boguraev. 2015. "Learning structures of negations from flat annotations."

Read, Jonathon, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2012. "Uio 1: Constituent-based discriminative ranking for negation resolution." *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 310–318.

Rosa, Rudolf. 2014. "Depfix, a tool for automatic rule-based post-editing of smt." *The Prague Bulletin of Mathematical Linguistics*, 102(1):47–56.

Szarvas, György, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. "The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts." *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 38–45.

Tamchyna, Aleš and Ondřej Bojar. 2015. "What a Transfer-Based System Brings to the Combination with PBMT." *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, 11–20. Stroudsburg, PA, USA.

Urešová, Zdeňka, Eva Fučíková, and Jana Šindlerová. 2016. "Czengvallex: a bilingual czech-english valency lexicon." *The Prague Bulletin of Mathematical Linguistics*. In prep.

Van der Plas, Lonneke, Paola Merlo, and James Henderson. 2011. "Scaling up automatic cross-lingual semantic role annotation." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 299–304.

Vilar, David, Jia Xu, Luis Fernando d'Haro, and Hermann Ney. 2006. "Error analysis of statistical machine translation output." *Proceedings of LREC*, 697–702.

Wetzel, Dominikus and Francis Bond. 2012. "Enriching parallel corpora for statistical machine translation with semantic negation rephrasing." *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 20–29.

Williams, Philip, Rico Sennrich, Maria Nadejde, Matthias Huck, and Philipp Koehn. 2015. "Edinburgh's syntax-based systems at wmt 2015." *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 199–209. Lisbon, Portugal.

Zhang, Hao, Licheng Fang, Peng Xu, and Xiaoyun Wu. 2011. "Binarized forest to string translation." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 835–845. Portland, Oregon, USA.

Zou, Bowei, Guodong Zhou, and Qiaoming Zhu. 2013. "Tree kernel-based negation and speculation scope detection with structured syntactic parse features." *EMNLP*, 968–976.