Applying the semantics of negation to SMT through n-best list re-ranking

Federico Fancellu
Centre for Global Intelligent Content
School of Computer Science and Statistics
Trinity College Dublin
ffancellu@cngl.ie

Bonnie Webber
School of Informatics
University of Edinburgh
Edinburgh, UK, EH8 9AB
bonnie@inf.ed.ac.uk

Abstract

Although the performance of SMT systems has improved over a range of different linguistic phenomena, negation has not yet received adequate treatment.
Previous works have considered the problem of translating negative data as one of data sparsity (Wetzel and Bond (2012)) or of structural differences between source and target language with respect to the placement of negation (Collins et al. (2005)). This work starts instead from the questions of *what is meant by negation* and *what makes a good translation of negation*. These questions have led us to explore the use of semantics of negation in SMT — specifically, identifying core semantic elements of negation (*cue, event* and *scope*) in a source-side dependency parse and re-ranking hypotheses on the n-best list produced after decoding according to the extent to which an hypothesis realises these elements.
The method shows considerable improvement over the baseline as measured by BLEU scores and Stanford's entailment-based MT evaluation metric (Padó et al. (2009)).

1 Introduction

Translating negation is a task that involves more than the correct rendering of a negation marker in the target sentence. For instance, translating *Italy did not defeat France in 1909* differs from translating *Italy defeated France in 1909*, or *France did not defeat Italy in 1909*, or *Italy did not conquer France in 1909*. These examples show that translating negation also involves placing in the right position the semantic arguments as well as the event directly negated. Moreover, if the source sentence was uttered in response to the statement *I think Italy defeated France in 1911*, where the focus is the temporal argument *in 1911*, one can see that the system should not lose track of the focus of negation when producing the hypothesis translation.

Although negation must be appropriately rendered to ensure correct representation of the semantics of the source sentence in the machine output, only some of the efforts to improve the translation of negation-bearing sentences in SMT address the problem.

Wetzel and Bond (2012) considered negation as a problem of data sparsity and so attempted to enrich the training data with negative paraphrases of positive sentences. Collins et al. (2005) and Li et al. (2009) both addressed differences in the placement of negation in source and target texts, by re-ordering negative elements in the source sentence to better resemble their position in the corresponding target text. Although these approaches show improvement over the baseline, neither considers negation as a linguistic phenomenon with specific characteristics.

This we do in the work presented here: We identify the elements of negation that an MT system has to reproduce and then devise a strategy to ensure that they are output correctly. These elements we take to be the *cue*, *event* and *scope* of negation[1]. Unlike previous works, we first validate the hypothesis that if the top-ranked translation in the n-best list does not replicate elements of negation from the source, there may be a more accurate translation after decoding, somewhere else on the n-best list. If the hypothesis is false, then problems in the translation of negation lie elsewhere.

[1] Due to its ambiguity and the fact that it is already included in the scope, we have ignored the *focus* of negation. That does not mean it may not be important to correctly reproduce the *focus*; there might be cases where, although not fully-capturing the *scope*, we want to translate correctly the part that is directly negated or emphasised.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 598–606,
Gothenburg, Sweden, April 26-30 2014.

We use dependency parsing as a basis for N-best list re-ranking. Dependencies between lexical elements appear to encode all elements of negation, offering a robust and easily-applicable way to extract negation-related information from a sentence. We carry out our exploration of N-best list re-ranking in two steps:

- First, an *oracle* translation is computed both to assess the validity of the approach and to understand the maximal extent to which it could possibly enhance performance. An oracle translation is obtained by performing n-best list re-ranking using reference translations as a gold-standard.
  To avoid the problem in Chinese-English Hierarchical Phrase-Based (HPB) translation of loss and/or misplacement of negation-related elements when hierarchical phrases are built, Chinese source sentences are first broken into sub-clauses Yang and Xue (2012), then translated and finally "stitched" back together for evaluation.

- Standard n-best list re-ranking is then performed using only *source-side* information. Hypotheses are re-ranked according to the degree of similarity between the negation-related elements in the hypotheses and those in the source sentence. Here the correspondence between source and target text is established through lexical translation probabilities output after training.

Results of this method show that n-best list reranking does lead to a significant improvement in BLEU score. However, BLEU says nothing about semantics, so we also evaluate the method using Stanford's entailment based MT metrics Padó et al. (2009), and also show improvement here. In the final section of the paper, we note the value of developing a custom metric that actually assesses the components of negation.

2 Related works

Negation has been a widely discussed topic outside the field of SMT, with recent works focused mainly on automatic detection of negation. Blanco and Moldovan (2011) have established the distribution of negative cues and the syntactic structures in which they appear in the WSJ section of the Penn Treebank, as a basis for automatically detecting scope and focus of negation using simple heuristics.

Machine-learning has been used by systems participating in the *SEM 2012 shared task on automatically detecting the scope and focus of negation. Those systems with the best F1 measures (Chowdhury and Mahbub (2012), Read et al. (2012) and Lapponi et al. (2012) all use a mixture of SVM (Support Vector Machines) and CRF. Their performance improves significantly when syntactic features are also considered. In particular, Lapponi et al. (2012) use features extracted from a dependency parse to guide their system to detect the correct scope boundary.

In translation, only few efforts have focussed on the problem of translating negation. Wetzel and Bond (2012) treat it as resulting from data sparsity. To remedy this, they enrich their Japanese-to-English training set with negative paraphrases of positive sentences, where negation is inserted as a 'handle' to the main verb after a sentence is parsed using MSR (Minimal Recursion Semantics Copestake et al. (2005)). Results show that BLEU score improves on a test sub-set containing *only negative sentences* when extra negative data is appended to the original training data and the language model is enriched as well. However, system performance deteriorates on both the original test set and on positive sentences. Moreover, generating paraphrases with negation expressed only on the main verb does not allow to fully capture the various ways negation can be expressed.

Other works considered negation in the framework of clause restructuring. Collins et al. (2005) pre-process the German source to resemble the structure of English while Li et al. (2009) tried to swap the order of the words in a Chinese sentence to resemble Korean. Rosa (2013) takes a post-processing approach to negation in English-Czech translation, "fixing" common errors such as the loss of a negation cue by either generating the morphologically negative form of the relevant verb (if the verb has such a form) or prefixing the verb with the negative prefix *ne*. Despite the improvements, these approaches do not really address what is special about negation.

3 Decomposing negation

Correctly translating negation involves more than placing a negative marker in the right position. We follow Blanco and Moldovan (2011) in decomposing negation into three main components:

- a *negation cue*, including negative markers, affixes and all the words or multiwords units that inherently express negation.
- a *negation event*, i.e. the event that is directly negated. Events can be either verbs (e.g. 'I do *not* **go** to the cinema) or adjectives (e.g. 'He is *not* **clever**').
- a *negation scope*, i.e. the part of the statement whose meaning is negated (Blanco and Moldovan, 2011, 229). The scope contains all those words that, if negated, would make the statement true. We follow here the guidelines for annotating negative data released during the *SEM 2012 Shared Task Morante et al. (2011) for a more detailed understanding on what to consider part of the negation scope.

In addition to these three components, formal semanticists identify a *negation focus*, i.e. the part of the scope that is directly negated or more emphasized. Focus is the most difficult part to detect since it is the most ambiguous. In the sentence 'he does not want to go to school by car' the speaker emphasized 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).

Translating negation is therefore a matter of ensuring that the *cue* is present, that its attachment to the corresponding *event* follows language-specific rules and that all the elements included in the scope are placed in the right order. Correctly reproducing the *focus* is left for future works.

4 Methodology

4.1 N-best list re-ranking

N-best list re-ranking is used in SMT to deal with sentence-level phenomena whose locality goes beyond n-grams or single hierarchical rules. It involves re-ranking the list of target-language hypotheses produced by decoding, using additional features extracted from the source sentence. In the case of negation, N-best list re-ranking allows us to assess whether a system is able to correctly translate the elements of negation, while failing to place the best hypothesis on these grounds at the top of the n-best list.

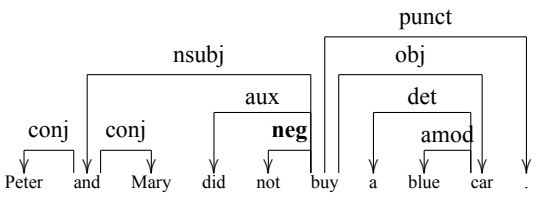
The current work follows the same approach as other n-best list re-rankers (Och et al. (2004); Specia et al. (2008); Apidianaki et al. (2012)) but using negation as the additional feature. Negation is here defined as the degree of overlap of *cue*, *event* and *scope* between the hypothesis translation and the source sentence.

Following Hasan et al. (2007), we use an n-best list of 10000 sentences but we do not initially tune the negation feature using MERT or interpolate it with other features. This is because in order to assess the degree of overlap between the scope in the source and the hypothesis sentence, a n-gram based score is used which conveys the same information as that of the *language model* score in the log-linear model. Moreover, our re-ranking exploits lexical translation probabilities, thereby resembling a simple *translation model*.

4.2 Extract negation using dependency parsing

The degree of overlap between the source sentence and the hypothesis translation is measured in terms of the overlap between their negation *cue*, *event* and *scope*. These must therefore be correctly extracted. Dependency parsing provides an efficient way to do so, with several advantages:

- Dependency parsing encodes the notions of *cue* and *event* as the dependant and the head respectively of a 'neg' relation. *Scope* can be approximated through *recursive retrieval of all the descendants of the verb-event*. The following example shows how these elements are extracted from the dependency parse:



nsubj(buy-6, and-2) , conj(and-2, Peter-1), conj (and-2, Mary-3), aux(buy-6, did-4), **neg(buy-6, not-5)**, root(ROOT-0, buy-6), det(car-9, a-7), amod(car-9, blue-8) , dobj(buy-6, car-9)

The 'neg' dependency relation conveys both the negation *cue* (**not-5**) and the negation *event* (**buy-6**) of the sentence 'Peter and Mary did not buy a blue car'. An approximate scope can be recovered by following the path from the event (included) to the terminal nodes and collecting all the lexical elements along the way.

Also in the case of a sentence containing a subordinate clause, dependency parsing is

able to correctly capture the latter as part of the *scope* given that the relative pronoun depends directly on the *event* of the main clause. On the other hand, recursion from the negated event excludes coordinate clauses that are *not* considered part of the scope, given that the event is a dependant of the connective.
One problem with this method is that it is unable to capture the entire scope when the head is nominal. For instance, *'no reasons were given'*, the 'neg' dependency holds between *no* and *reasons* but it needs to climb the hierarchy further to get to the verbal head *given*. The same holds for negation on object nominals. We leave this to future work (along with affix-conveyed negation), needing to show first that the current approach is a good one.

- A dependency parser can be developed for any language for which a Treebank or Propbank is available for training. This extends the range of source languages to which the approach can be applied.

4.3 Computing an oracle translation

In order to test the validity of the method and to assess its maximum contribution, we first use it with an *oracle translation* in which n-best list reranking relies on a comparison with negation *cue*, *event* and *scope* extracted from the reference translation(s), here assumed to correctly contain all elements of negation.

Each hypothesis on the n-best list is assigned an overlap score with these reference-translation-derived elements, and the hypothesis with the highest score is re-ranked at the top and used for evaluation.

The overlap score is obtained by summing up three sub-scores: (i) the *cue overlap* score measures how many cues in reference are represented in the hypothesis, normalised by the number of cues in the reference; (ii) the *event overlap* score measures how many events in the reference are represented in the hypothesis, normalised by the number of the events in the reference; and (iii) the *scope overlap* score is a weighted n-gram overlap between hypothesis scope and reference scope, with higher weight for higher-order n-grams. Given less-than-perfect machine output, breaking down the score into subscores allows us to consider different degrees of correctness in translating negation. When multiple reference translations are available, the hypothesis is matched with each, and only the best score taken into consideration.

4.4 Re-ranking using lexical translation probabilities

After the oracle translation is computed, traditional n-best list re-ranking is performed relying on *source side information only*. We then bridge the gap between source and target language using *lexical translation probabilities* to render source-side *cue*, *event*, *scope* into the target language. Re-ranking involves three separate steps:

- The source sentence is parsed and dependencies extracted. Since the present work tackles Chinese-to-English translation, we had to enhance the representation of negative dependencies in the Chinese source, where only the adverb 不 *bu4* is flagged as 'neg' dependant. To do this, we follow the same intuition used to isolate negation-bearing sentences in the test set (see section 5).

- To obtain a rough translation of *cue*, *event* and *scope* in the target language, the top ten lexical translation probabilities for each lexical item, available in the lexical translations (in order of probability) table output after training, are considered.

- Hypotheses in the n-best list are re-scored taking into consideration the information above. Scoring *cue* and *event* is straightforward; the words for the *cue* and the *event* are assigned the lexical probability of being the translation of the *cue* and the *event* in the Chinese sentence by looking up the lexical translation table. If the *cue* or the *event* do not figure as translations of the negation *cue* and *event* in the Chinese sentence, a score of 0 is assigned to them.
The scope is instead scored by looping through the words in the hypothesis; for each such hypothesis word, the process identifies which source-side scope word it is most likely to be the translation of. If no *scope* can be retrieved, a score of 0 is given for scope matching. For each word the best translation probability is taken into account and these are then summed together to score how likely is the

scope in the hypothesis to be the translation of the scope in the source.

5 System

A hierarchical phrase based model (HPBM) was trained on a corpus of 625000 (∼ 18.200.000 tokens) length-ratio filtered sentences. 56949 sentences (∼ 9.11%) of the Chinese side and 48941 sentences (∼ 7.83%) of the English side of the training set were found to include at least one instance of negation. 2500 sentences were instead included in the dev set to tune the log-linear model using the MERT algorithm. 3725 sentences from the Multiple-Translation Chinese Corpus 2.0 (LDC2002T01) were used as test set. The test set comes divided into four sub-sets; in this paper these sub-sets are referred as test set 1 to 4. The source side was tokenized using the Stanford Chinese Word Segmenter (Tseng et al. (2005)) and encoded in 'utf-8' so to serve as input to the system. In order to focus on the problem of translating negative data, the 563 sentence pairs containing negation were extracted from the original test set. This test set constitutes the true baseline improvements will be measured upon. Reducing the number of test sentences also eases the computation load when involved in dependency parsing on 10.000 sentences in each n-best list. Negated sentence were isolated by means of both regular expressions and dependency parsing; this is because, as pointed out above, the Chinese side does not flag all negative dependencies as such.[2]

6 Results

6.1 Baseline

BLEU scores for the baseline systems are given in Table 1, where the *negative* subset is compared to the *original* (all sentences) and *only positive sentences* conditions.

Table 2 shows instead the result for the negative baseline across three different metrics. Along with the BLEU scores, we also took into consideration an entailment-based MT evaluation metric, the RTE score[3] Padó et al. (2009). The RTE score assesses to what extent the hypothesis entails the reference translation across a wide variety of semantic and syntactic features. Another reason we chose this metric is because it contains a feature for polarity as well as features to check the degree of similarity between the syntactic tree and the dependencies between hypothesis and reference translation, the latter being what we used to recover the elements of negation. We expect this metric to give a further insight on the quality of the machine output.

Baseline results are in line with the results of Wetzel and Bond (2012), where there is a drop in BLEU scores between positive and negative sentences, and between the overall test set and the one containing negative data only.

When analysing the results from the baseline, we noticed that words were being deleted or moved inappropriately when the hierarchy of phrases was being built. This might be detrimental to the translation of negation since elements might end up outside the correct negation scope. The following example illustrates this problem.

(1) *Source*：三 年 来， 这些 城市 累计 完成 固定 资产 投资 一百二十亿 元， 昔日 边境 城市 的 “楼 不 高， 路 不 平、灯 不 明、水 不 清、通讯 不 畅” 的 状 况 **已 得到 了 改变** 。

Baseline : Investment in fixed assets investment in the three years , 一百二十亿 yuan , **floor is not high** , ” the former border city , road , and **communication conditions have not been completed** , **will not change .**

Due to unrestricted rule application, mainly guided by the language model, the underlined clauses containing negation on the source side have been deleted. Moreover, the polarity of the last clause, positive in the source, is changed into negative in the target translation, most probably because a negative cue is moved from somewhere else in the sentence.

In order to solve these two problems, we exploit the syntactic feature of the Chinese language of grouping clauses into a single sentence. We follow the intuition of Yang and Xue (2012) in using

[2] While the English dependency parser is able to identify almost all negative markers and their dependencies, the Chinese dependency parser here deployed (the Stanford Chinese Dependency parser) only captures sentences containing the adverb 不*bu4*. For this reason, we exploited the list of negation adverbs included in the Chinese Propbank (LDC2005T23) documentation and look for each of them via regular expressions. Moreover we also looked for words containing 不 as component since they are most likely to carry negative meaning (e.g. 不久, 'not-long').

[3] The entailment-based MT metric also outputs an RTE+MT score, where the RTE score is interpolated with traditional MT metrics (e.g. BLEU, NIST, TER, METEOR).

Test set	Original	Positive	Negative	Orig. → Neg.	Pos. → Neg.
Test 1	32.92	**32.95**	29.64	- 3.28	- 3.31
Test 2	25.88	**26.21**	24.31	- 1.57	- 1.9
Test 3	19.00	**19.78**	16.11	- 2.89	- 3.67
Test 4	28.64	**29.71**	27.14	- 1.5	- 2.57
Average	26.61	**27.16**	24.3	-2.31	- 2.96

Table 1: BLEU scores for the baseline system. The difference in BLEU scores between the *positive*, the *original* and the *negative* conditions is also reported.

Test set	BLEU	RTE	RTE+MT
Test 1	29.64	0.22	0.837
Test 2	24.31	0.307	0.732
Test 3	16.11	-0.603	-0.095
Test 4	27.14	-0.25	0.33
Average	24.3	-0.08	0.451

Table 2: BLEU, RTE and RTE+MT scores for the baseline system as tested on the sub-set only containing negative sentences.

commas to guide the segmentation of a sentence into constituent sub-clauses. Moreover, we also use other syntactic clues to segment the test sentences, including quotes in direct quotation, to reduce the size of the test sentences.
The constituent sub-clauses are then translated singularly and 'stitched' back together into the original sentence for evaluation.

6.2 Re-ranking results

Table 3 and 4 shows the performance of the system when n-best list re-ranking is performed. Table 3 shows the results for the oracle translation, while Table 4 the results for actual n-best list re-ranking. Two conditions are here compared: a *short* condition where test sentences are chunked into constituent sub-clauses prior to translation and a *original (orig.)* condition where no chunking is performed.

Results shows considerable improvements over the baseline when re-ranking is performed — an average BLEU score improvement of 1.75 points. As hypothesised, we get further improvement when Chinese source sentences are translated through their constituent sub-clauses — an average BLEU score improvement of 3.07 points. A similar improvement is shown in Table 5 where the original test sets comprising both positive and negative sentences are considered. This proves the validity of n-best list re-ranking using syntactic dependencies as a method to improve the quality of the translation of negative data. The following example shows the improvement in detail:

(2) *Source* : 关于 通货膨胀 ， 伊辛 说 ， 欧元区 通涨率 整体 呈 下降 态势 ,目前 没有 迹象 表明 在 经济 发展 的 中期 阶段 将 出现 物价 不 稳定 的 风险
Ex. reference : When asked about inflation, he said : "The overall inflation rate in the Euro area still exhibits a down trend. At present, there is no sign to show economic development in the medium term will create risks of price instability".
Baseline : on the inflation he said the euro dropped overall medium term economic development will in no signs of inflation risks .
Oracle : on inflation , said the euro dropped overall there is no signs of economic development in the medium term prices will not risks .
Source-only re-ranking : on inflation , said the euro dropped overall there is no signs of economic development in the medium term will price risks .

In (2) the baseline translation shows the problems mentioned earlier, where movement leaves negation with the wrong scope, changing the overall meaning of the sentence. Decomposing sentences into constituent clauses and then re-ranking the translations permits negation to retain its correct scope so that the meaning is the same as the reference sentence.

7 Conclusion

We have presented an approach to translating negative sentences that is based on the semantics of negation and applying it to n-best list re-ranking.

		BLEU	RTE	RTE+MT
1	Baseline	29.64	0.22	0.837
	Orig.	33.73 (+4.09)	0.64 (+0.42)	1.396 (+0.559)
	Short	**35.39 (+5.75)**	**0.74 (+ 0.52)**	**1.508 (+ 0.671)**
2	Baseline	24.31	0.307	0.732
	Orig.	27.43 (+3.12)	0.457 (+0.15)	1.12 (+0.388)
	Short	**27.29 (+3.18)**	**0.6 (+ 0.293)**	**1.175 (+ 0.443)**
3	Baseline	16.11	-0.603	-0.095
	Orig.	17.97 (+1.86)	**0.356 (+ 0.959)**	**0.958 (+ 1.053)**
	Short	**18.19 (+2.08)**	0.243 (+ 0.84)	0.78 (+ 0.875)
4	Baseline	27.14	-0.25	0.33
	Orig.	31.97 (+ 4.83)	0.42 (+ 0.67)	1.024 (+ 0.694)
	Short	**32.50 (+ 5.36)**	**0.57 (+ 0.82)**	**1.36 (+ 1.03)**
Avg.	Baseline	24.3	- 0.08	0.45
	Orig.	27.78 (+ 3.48)	0.47 (+ 0.55)	1.12 (+ 0.67)
	Short	**29.09 (+ 4.79)**	**0.52 (+ 0.60)**	**1.23 (+ 0.78)**

Table 3: BLEU, RTE and RTE+MT scores for the oracle translation. The test sets evaluated are marked from 1 to 4. Improvement over the baseline is reported.

		BLEU	RTE	RTE+MT
1	Baseline	29.64	0.22	0.837
	Orig.	31.96 (+ 2.32)	0.62 (+ 0.4)	1.382 (+ 0.545)
	Short	**34.20 (+ 4.56)**	**0.68 (+0.46)**	**1.452 (+ 0.615)**
2	Baseline	24.31	0.307	0.732
	Orig.	26.65 (+2.34)	0.48 (+ 0.173)	1.159 (+ 0.427)
	Short	**26.94 (+ 2.63)**	**0.49 (+0.183)**	**1.172 (+ 0.44)**
3	Baseline	16.11	-0.603	-0.095
	Orig.	17.20 (+ 1.09)	**0.35 (+ 0.953)**	**0.935 (+ 1.03)**
	Short	**17.41 (+ 1.3)**	0.226 (+0.829)	0.87 (+ 0.965)
4	Baseline	27.14	-0.25	0.33
	Orig.	28.42 (+ 1.28)	0.302 (+ 0.552)	1.01 (+ 0.68)
	Short	**30.96 (+ 3.82)**	**0.55 (+ 0.8)**	**1.36 (+ 1.03)**
Avg.	Baseline	24.3	-0.08	0.45
	Orig.	26.05 (+ 1.75)	0.438 (+ 0.518)	1.12 (+ 0.669)
	Short	**27.37 (+ 3.07)**	**0.51 (+ 0.59)**	**1.21 (+ 0.759)**

Table 4: BLEU, RTE and RTE+MT scores for the sentences re-ranked using source side information only. Improvement over the baseline is reported.

Dependency parsing and lexical translations are here considered as easily applicable methods to extract and translate negation related information across different language pairs. Improvements across different automatic evaluation metrics show that the above method is useful when translating negative data. In particular, the entailment-based RTE metric is here used as an alternative to the BLEU score given the semantic and syntactic features assessed, polarity included. Given the positive results, one can conclude that the problem is neither one of data sparsity nor syntactic mismatch.

We have also demonstrated that when dealing with sentences containing multiple sub-clauses, translating the constituent sub-clauses separately and then stitching them back together before evaluation avoids the loss or excessive movement of negation during decoding. This was evident in the case of Chinese and HPBMs but there is no reason why this does not hold also for other languages.

8 Future works

Given the validity of the present approach, future works should be focused in extending it to different language pairs. Also, it would be useful to research more in detail into language typology and try to devise a method which is language independent.

Although leading to an overall improvement, n-best list re-ranking does not always guarantee a perfect translation. It is therefore useful in the future to investigate ways of always ensuring that the n-best list contains a good translation of negation by, for instance, enriching the hypotheses list with paraphrases. Post-editing rules can also be considered to further correct the final output.

Finally, although we can show considerable improvement with respect to both n-gram overlap

		BLEU	RTE	RTE+MT
1	Baseline	32.92	-0.49	-0.073
	Orig.	33.54 (+ 0.62)	-0.38 (+ 0.11)	0.046 (+ 0.119)
	Short	**34.02 (+ 1.1)**	**-0.33 (+ 0.16)**	**0.057 (+ 0.13)**
2	Baseline	25.88	-2.173	-1.726
	Orig.	26.3 (+ 0.42)	-1.851 (+ 0.322)	-1.376 (+ 0.35)
	Short	**26.42 (+ 0.54)**	**-1.80 (+ 0.373)**	**-1.339 (+ 0.387)**
3	Baseline	19.00	-0.897	-0.644
	Orig.	19.20 (+ 0.20)	**-0.731 (+ 0.166)**	**-0.474 (+ 0.17)**
	Short	**19.23 (+ 0.23)**	-0.743 (+ 0.154)	-0.488 (+ 0.156)
4	Baseline	28.64	-3.43	-3.16
	Orig.	29.56 (+ 0.92)	-3.01 (+ 0.42)	-2.72 (+ 0.44)
	Short	**29.95 (+ 1.31)**	**-2.94 (+ 0.49)**	**-2.67 (+ 0.49)**
Avg.	Baseline	26.61	-1.747	-1.4
	Orig.	27.15(+ 0.54)	-1.488 (+ 0.259)	-1.131 (+ 0.269)
	Short	**27.41(+ 0.8)**	**-1.453 (+ 0.294)**	**-1.11 (+ 0.29)**

Table 5: BLEU, RTE and RTE+MT scores for the the original test set, containing both positive and negative sentences re-ranked using source side information only. Improvement over the baseline is reported.

with the reference translation (BLEU score) and overall semantic similarity, it remains to be determined the extent to which the machine output captures elements of negation present in the reference translation and on which system improvement depends. A more targeted metric is needed, that can effectively determine the extent to which cue, event and scope are captured in hypothesis translation as compared to the reference gold standard. That is the subject of current and future work (Fancellu (2013)), which should implement the new customized metric to include measures of precision, recall and a F1 measure.

References

Apidianaki, M., Wisniewski, G., Sokolov, A., Max, A., and Yvon, F. (2012). Wsd for n-best reranking and local language modeling in smt. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 1–9. Association for Computational Linguistics.

Blanco, E. and Moldovan, D. I. (2011). Some issues on detecting negation from text. In *FLAIRS Conference*.

Chowdhury, M. and Mahbub, F. (2012). Fbk: Exploiting phrasal and contextual clues for negation scope detection. In *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*, pages 340–346. Association for Computational Linguistics.

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 531–540. Association for Computational Linguistics.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (2005). Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2-3):281–332.

Fancellu, F. (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.

Hasan, S., Zens, R., and Ney, H. (2007). Are very large n-best lists useful for smt? In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 57–60. Association for Computational Linguistics.

Lapponi, E., Velldal, E., Øvrelid, L., and Read, J. (2012). Uio 2: sequence-labeling negation using dependency features. In *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*, pages 319–327. Association for Computational Linguistics.

Li, J.-J., Kim, J., Kim, D.-I., and Lee, J.-H. (2009). Chinese syntactic reordering for adequate generation of korean verbal phrases in chinese-to-

korean smt. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 190–196. Association for Computational Linguistics.

Morante, R., Schrauwen, S., and Daelemans, W. (2011). Annotation of negation cues and their scope: Guidelines v1. Technical report, 0. Technical report, University of Antwerp. CLIPS: Computational Linguistics & Psycholinguistics technical report series.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., et al. (2004). A smorgasbord of features for statistical machine translation. In *HLT-NAACL*, pages 161–168.

Padó, S., Galley, M., Jurafsky, D., and Manning, C. D. (2009). Textual entailment features for machine translation evaluation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 37–41. Association for Computational Linguistics.

Read, J., Velldal, E., Øvrelid, L., and Oepen, S. (2012). Uio 1: Constituent-based discriminative ranking for negation resolution. In *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*, pages 310–318. Association for Computational Linguistics.

Rosa, R. (2013). Automatic post-editing of phrase-based machine translation outputs. Master's thesis, Institute of Formal and Applied Linguistics, Charles University, Prague.

Specia, L., Sankaran, B., and Nunes, M. d. G. V. (2008). N-best reranking for the efficient integration of word sense disambiguation and statistical machine translation. In *Computational Linguistics and Intelligent Text Processing*, pages 399–410. Springer.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. (2005). A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 171. Jeju Island, Korea.

Wetzel, D. and Bond, F. (2012). Enriching parallel corpora for statistical machine translation with semantic negation rephrasing. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 20–29. Association for Computational Linguistics.

Yang, Y. and Xue, N. (2012). Chinese comma disambiguation for discourse analysis. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 786–794. Association for Computational Linguistics.