Translating Negation: A Manual Error Analysis

Federico Fancellu and Bonnie Webber
School of Informatics
University of Edinburgh
11 Crichton Street, Edinburgh
f.fancellu[at]sms.ed.ac.uk,bonnie[at]inf.ed.ac.uk

Abstract

Statistical Machine Translation has come a long way improving the translation quality of a range of different linguistic phenomena. With negation however, 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. In contrast, the current paper shows that an informative empirical error analysis can be formulated in terms of (1) the set of semantic elements involved in the meaning of negation, and (2) a small set of string-based operations that can characterise errors in the translation of those elements. Results on a Chinese-to-English translation task confirm the robustness of our analysis cross-linguistically and the basic assumptions can inform an automated investigation into the causes of translation errors. Conclusions drawn from this analysis should guide future work on improving the translation of negative sentences.

1 Introduction

In recent years, there has been increasing interest in improving the quality of SMT systems over a wide range of linguistic phenomena, including coreference resolution (Hardmeier et al., 2014) and modality (Baker et al., 2012). Amongst these, however, translating negation is still a problem that has not been researched thoroughly.

This paper takes an empirical approach towards understanding why negation is a problem in SMT. More specifically, we try to answer two main questions:

1. *What kind* of errors are involved in translating negation?
2. What are the causes of these errors during decoding?

While previous work (section 2) has shown that translating negation is a problem, it has not addressed either of these questions.

The present paper focuses on the first one; we show 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 (section 3). We report the results of the analysis of a Hierarchical Phrase Based Model (Chiang, 2007) on a Chinese-to-English translation task (section 4), where we show that all error categories occur to some extent with scope reordering being the most frequent (section 5).

Addressing question (2) requires connecting the assumptions behind this manual error analysis to errors occurring along the translation pipeline. As such, we complete the analysis by briefly introduce an automatic method to investigate the causes of the errors at decoding time (section 6).

Conclusion and future works are reported in section 7 and 8.

2 Previous Work

In recent years, automatic recognition of negation has been the focus of considerable work. Following Blanco and Moldovan (2011) and Morante and Blanco (2012) detecting negation is a task of unraveling its structure, i.e. locating in a text its four main components:

- **Cue**: the word or multi-word unit inherently expressing negation (e.g. 'He is not driving a car')
- **Event**: the lexical element the cue directly refers to (e.g. 'He is not driving a car')
- **Scope**: all the elements whose falsity would prove negation to be false; given that the cue is not included, the scope is often discontinuous (e.g. 'He is not driving a car')
- **Focus**: the portion of the statement negation primarily refers to (e.g. 'He is not driving a car).

The *SEM 2012 shared task represented a first attempt to apply machine learning methods to the problem of automatically detect the aforementioned elements in English. In particular CRFs and SVMs, making use of syntactic (both constituent and dependency based) clues, were shown to lead to the best results in a supervised machine learning setting (Read et al., 2012; Chowdhury and Mahbub, 2012). The shared task also saw the release of a fully annotated corpus in the literature domain, which represents, along with the BioScope corpus (Szarvas et al., 2008), the only resource specifically annotated for negation.

There were also a few attempts in automatically detecting negation in Chinese texts. Li et al. (2008) designed a negation detection algorithm based on syntactic patterns; similarly, Zheng 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.

In a bilingual setting such as the SMT, however, most work has only considered negation as a side problem. For this reason, no actual analysis on the type of errors involved in translating negation or their causes has been specifically carried out. The standard approach has been to formulate an hypothesis about what can go wrong when translating negation, modify the SMT system in a way aimed at reducing the number of times that happens, and then assume that any increase in BLEU score - the standard automatic evaluation metric used in SMT - confirms the initial hypothesis. Collins et al. (2005) and Li et al. (2009) consider negation, along with other linguistic phenomena, as a problem of *structural mismatch* between source and target; Wetzel and Bond (2012) consider it instead as a problem of *training data sparsity*; finally Baker et al. (2012) and Fancellu and Webber (2014) consider it as a *model problem*, where the system needs enhancement with respect to the semantics of negation. Given that all these works assess the quality of translation of negative sentences using an *n-gram* overlap metric, there is no certainty whether any improvement derives from a better rendering of negation or from other, non-negation related elements.

Evaluating the semantic adequacy of the SMT output has also stimulated interest in recent years. Traditional error categories, such as the ones presented in (Vilar et al., 2006), are mostly based on n-gram overlap between hypothesis and reference and so are the most widely used automatic evaluation metrics used in SMT (e.g. BLEU (Papineni et al., 2002) and TER (Snover et al., 2009)). In contrast, MEANT (Lo and Wu, 2010, 2011) and its human counterpart, HMEANT, attempt to abstract from simple string matching and assess the degree of semantic similarity between machine output and reference sentence. To do so, both sides are annotated using Propbank-like semantic labels, and the fillers matched if both sides contain the same event. To assign a score to the test set evaluated, an F_1 measure over precision and recall of matched fillers is then computed.

3 Methodology

3.1 Manual Annotation

First, we start with the assumption that negation is a language independent semantic phenomenon which can be defined as a structure. This assumption implies that it should be possible to annotate any language using the elements in this structure – **cue**,

event and **scope**. Isolating a small set of semantic elements involved in the construction of negation is useful in the context of SMT to reduce negation into tangible elements at the string level. Moreover each of the three elements above represents different translation problems: if, for instance, translating the *cue* mainly involves ensuring the presence of a negation marker, translating the *scope* involves instead ensuring that semantic elements are translated in the right domain and most of the times, around the negated *event*.

We carried out the annotation of *cue*, *event* and *scope* on both the source Chinese sentences and the correspondent translation output by the SMT system, following the guidelines released during the SEM* 2012 shared task (Morante et al., 2011). To our understanding, this is the first work that applies these guidelines to a language other than English. It is however worth noticing 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 negation scoping on non-factual events, 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. partially correct event (cf. section 3.2).

3. For the same reason as (2), the event in a nominal predicate includes all its modifiers.

4. All these points apply to Chinese as well; in addition, 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*.

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 fillers, here taken to be Propbank-like semantic arguments. In doing 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.

Example (1) illustrates our annotation scheme over the first instance of *bù* (not) in a Chinese source sentence.

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

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

As shown in (1) 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. We instead consider the second instance of *bù/not* as 'non-functional negation' and do not annotate it since it is just part of the question and does not constitute itself a negation instance.

3.2 Manual Error Analysis

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 consider three main operations 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).

- **Insertion**: the negation element is not present in the source sentence but has been inserted in the machine output. This resembles the *extra words* sub-category in the *incorrect words* class.

- **Reordering**: whether the element has been moved outside its scope. Since some semantic elements can also move inside the scope and take a role which they did not have in the original source sentence, we define the former reordering error as *out-of-scope* reordering error and the latter *intra-scope* reordering error. The reordering category represents an adaptation of the original *word order* category.

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 (a.k.a. the *decoding* 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 Chinese word or sequence of Chinese words (generally referred in SMT as *phrases*) has not been seen during training (so called *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 translated unit is *correct* or *partially correct*. We also consider *correct* translations that are synonyms of the source negation element 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). We consider *partially correct* translated elements that do not contain errors which impact the overall meaning. 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 $e \in$ E = {cue,event,filler}. However, unlike HMEANT, we do not normalise the number of correct fillers by the number of total fillers in the predicate.

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

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

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

4 System

We carried out the error analysis on the output of the Chinese-to-English hierarchical phrase based system submitted by the University of Edinburgh for the NIST12 MT evaluation campaign.

Hierarchical phrase-based (or HPB) systems are a class of SMT systems that use syntax-like rules and hierarchical tree structures to build an hypothesis translation given a test source sentence and a model previously trained on a bilingual corpora. Unlike pure syntax models, HPBMs do not make use of syntactic constituent tags for non-terminals but instead use an X as placeholder for recursion. A rule used in a Hierarchical Phrase based system looks like the following,

ne veux plus X_1 → do not want X_1 anymore

where the French source (also referred to as the *left hand side* - LHS of the rule) and the English target side (the *right hand side* - RHS) allows arbitrary insertion of another rule where the placeholder X is located.

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 inspirational 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.

In order to carry out future work on the effect of word segmentation on the elements on negation, we built two different systems (and therefore *rule tables*), one from data segmented using the LDC-WordSegmenter and the other using the Stanford Word Segmenter. The former matches the segmentation of the NIST08 test set, whilst the latter the one of the IWSLT14 test set.

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.

5 Results

5.1 Manual Analysis

5.1.1 NIST MT08

The results of the manual evaluation for the NIST MT08 test set are reported in Table 1. 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 of cues 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 reordered 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 (2) and (3) respectively.

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

Ref: He never showed any bias against me [because i 'd complained to him]$_{sub}$ [...]

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

(3) [...] *jiù huì rènwéi bù cúnzài*
[...] then can think not exist

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

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

As for the translation of events, a trend similar to the translation of cues can be observed, although the percentage of deletions is higher than the cue. The trace shows that in 3 out of 11 cases, deletion is caused by an OOV item, i.e. a Chinese phrase which is not seen in training and for which the system has not learned any translation. The remaining cases resemble the cue case, insofar as no rule contains the target side event. Another problem arising with events is that some fillers in the source might have erroneously become events in the machine output and vice versa; we found 3 events on the source becoming fillers in the target and 7 fillers on the source becoming events in the machine output, as shown in (4).

(4) *zhè yīge jiēduàn de biăxiàn shì* [*duănqī*
This one stage of show is short-term
xiāoguō]$_{filler}$ *bùdà*$_{cue+event}$ [...]
result not big [...]

Ref: what this stage brings forward is : modest success in the short-term [...]

Hyp: this is a stage performance are not$_{cue}$ [short-term effect]$_{event}$

The fact that most of reordering errors are filler-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 fillers should be attached to and in what order, *in-scope* and *out-of-scope* problems are to be expected.

Around 10% of filler-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 contain the filler on the target side.

Finally, in the case of fillers, we notice that 2 of the incorrect fillers in the hypothesis were due to the *insertion* in the scope of fillers 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 (5) a rule that translates *yĭzhìyú* ('to the extent of') into 'to the extent of *they*' is used, adding a filler to following negation scope.

(5) [...] *yĭzhìyú wúfă yú oū zhou*
[...] to the extent not possible with Europe
méngguó zhèngcháng zhănkāi hézuò
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]$_{filler}$ are unable to conduct normal with its european allies cooperation

NIST MT08 test set					
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		
Fillers	98		80		
	#	R%	#	P%	F_1
Correct cues	58/66	87.87	53/57	92.98	90.35
Correct events	34/66	51.51	29/57	50.88	
+ Partial events	34 + **8**/66	57.6	29 + **8**/57	57.9	57.74
Correct fillers	48/98	48.97	45/80	56.25	
+ Partial fillers	48 + **9**/98	58.16	45 + **9**/80	67.5	62.48
Deleted cues	4/66	6			
Deleted events	11/66	16.6			
Deleted fillers	9/98	9.18			
Inserted fillers			2/80	2.5	
Reordered cues *same scope*	1/66	1.5	1/57	1.75	
Reordered cues *out of scope*	3/66	4.5			
Reordered events *same scope*	3/66	4.5	7/57	12.2	
Reordered events *out of scope*	1/66	1.5			
Reordered fillers *same scope*	8/98	8.16	5/80	6.25	
Reordered fillers *out of scope*	21/98	21.41			

Table 1: Results from the error analysis of the 250 sentences randomly extracted from the NIST MT08 test set.

5.1.2 IWSLT '14 Tst2012 TED Talks

Results for the TED talks test set are reported in Table 2. 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 fillers on the overall number translation errors might be connected to the fact that on average sentences in the TED talk, also given their domain, are shorter than the sentences in the NIST08 test set and therefore there is less chance of operating long range reordering.

We can also observe that genre has an effect on the number of 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, 3 of event deletion and 5 of filler 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 *zĕnme*, which end up being collapsed in a single word unit.

(6) *bùzhīzěnme* *yòng wǒmen bù néng*
do not know how use we not be able
wánquán lǐjiě *de fāngshi* [...]
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 (5), it might have been that a rule containing extra material was preferred because a better fit in that specific context (a LM score is in fact 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 Towards An Automatic Error Analysis

This manual error analysis assesses the quality of the 1-best translation output by the system. More can be done: (i) we can determine which component of the system is responsible for each error so as to know where to intervene and (ii) we can automate the whole process of error finding. Both actions can be referred to as *automatic error analysis*, given that they rely on (semi-)automatic method to analyse errors in translating negation, although they differ in the scope of their analysis: (i) represents an extension of the manual error analysis, whilst (ii) aims at automating it.

IWSLT14 tst2012 TED talks					
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		
Fillers	103		83		
	#	R%	#	P%	F_1
Correct cues	61/69	88.4	53/54	98	92.95
Correct events	48/69	69.56	40/52	76.92	
+ Partial events	48 + **3**/69	71.73	40 + **3**/52	79.8	75.55
Correct fillers	64/103	62	64/83	77	
+ Partial fillers	64 + **3**/103	63.59	64 + **3** /83	78.9	70.42
Deleted cues	7/69	10.14			
Deleted events	5/69	7.2			
Deleted fillers	4/103	3.8			
Inserted cue			1/54	1.8	
Inserted fillers			1/83	1.2	
Reordered events *same scope*	5/69	7.2	1/52	1.9	
Reordered events *out of scope*	4/69	5.7			
Reordered fillers *same scope*	2/103	1.9	6/83	7.2	
Reordered fillers *out of scope*	13/103	12.62			

Table 2: Results from the error analysis of the 250 sentences randomly extracted from the IWSLT2014 test set.

Although out of the scope of the present work, we briefly sketch our current work on (i) whilst leaving (ii) for future work. The reason for this is because the assumption behind this as many other manual analysis, i.e. that a small set of string-based error categories can be used to characterise different kind of translation errors (here semantic), can be easily projected in the automatic error analysis. Moreover, the importance and indispensability of a manual error analysis is highlighted when devising an automatic error analysis. This is obvious in the case of (i), where, in order to find the causes of the errors, we need to know what these are. However, even we succeed in (ii) and we are able to spot errors automatically, we still need a manual error analysis as a benchmark to assess the quality of any automatic method.

When we talk about detecting errors during decoding, we try to determine the reason why our system is behaving differently to what we expect. These expectations depends on the source side negation element processed at each step during decoding and are closely linked to both the set of string-based error category and the set of negation sub-constituents used in this manual error analysis:

1. The **cue** has to be translated correctly; no cue **deletion** or **insertion** should occur.

2. The **event** has to be translated correctly; no event **deletion** or **insertion** should occur.

3. The **cue** has to be connected to right **event**; no cue or event **reordering** should occur.

4. The **semantic arguments** in the source scope should be translated and reproduced in line with the target language semantics; no **deletion**, **insertion** or **reordering** of the semantic fillers should occur.

An ideal system would meet all the above conditions in translating negation in each cell of the decoding chart[1]; if not, we have to inspect the decoding chart trace and classify the errors occurred. The goal here is to find which part of the translation system is responsible for each error category. There are three main type of errors, each one connected to one component of the translation pipeline:

- **Induction errors**, where the correct translation for a given element is absent from the search space. These errors depends on how many target translations are fetched from the rule table when a given source span is translated (default is 20). The more we consider, the more likely is for the correct translation to be inserted in the search space. The system component related to this category is the **rule table**.

- **Search errors**, where the correct translation fetched from the rule table disappears from the search space before making it to the final cell, due to pruning or other optimisation heuristics. The system component responsible for this error is the **search space**.

- **Model errors**, where the system ranks bad translations better than better translations. The component responsible is the **scoring function**.

Induction errors occur when no negation element is found in any hypothesis built in any of the chart cell; search errors occur when, by enlarging the search space, hypotheses meeting the expected conditions that were previously absent from the chart are now present; finally, model errors occur where hypothesis meeting one or more expectation are present but rank lower than the ones that do not.

Since these expectations are based on source side elements we need a way to project source side negation elements into a target language; for expectation (1) and (2), we are experimenting with two different methods: (i) via an automatically extracted list of potential cues and a bilingual dictionary enriched with paraphrases; (ii) by extracting cues and events from the multiple reference translations set. To ensure that expectation (3) and (4) are met we instead use a dependency parse.

Preliminary results on the translation of the cue alone (expectation (1)) in the NIST08 MT test set shows that it is uniquely a problem of model error, where good hypotheses are ranked lower than bad ones. A comparison between the scores of good and bad hypotheses show, when the former are not ranked properly, shows that it is the translation model the main responsible for such bad ranking.

7 Conclusion

The present paper presents an analysis of the errors involved in translating negation. We showed that it is possible to build a clear and robust error analysis using (1) the set of semantic elements involved in the meaning of negation (**cue**, **event** and **scope**) and (2) a sub-set of string-based operations traditionally used in SMT error analysis (**deletion**, **insertion** and **reordering**).

Results of a manual error analysis on a Chinese-to-English output shows that this analysis is easily portable to a language other than English and allows us to cover a wide range of potential errors occurring during translating. Our findings also show that amongst the three elements of negation here considered, the *scope* is the most problematic and reordering is in general the most frequent error in Chinese-to-English translation. In the case of deletion or insertion of negation elements, we also found that the errors are attributable to a rule application that

[1] Hierarchical phrase-based decoder uses a variant of bottom-up CKY chart algorithm

prefers positive translations over negative and are therefore not caused by OOV items not seen during training.

Using the assumptions and the results of the manual error analysis, we also introduced an automatic way to inspect the causes of the errors in the decoding chart trace. Preliminary results show that the scoring function is the main responsible for cue deletion errors observed.

We hope that the methodology and the results of the present work can guide future work on improving the translation of negative sentences.

8 Future Work

In the present paper, we have successfully applied the manual error analysis to the output of a Chinese-to-English Hierarchical Phrase-based system. Future work will extend this method to other language pairs and different SMT systems. We in fact expect these two variables to impact the kind of errors found in translation. Chinese and English are in fact very similar in the way they express negation: adverbial negation is the most frequent way of expressing negation (Blanco and Moldovan, 2011; Fancellu and Webber, 2012); morphological negation (or *affixal*) or lexically embedded negation is present in both languages and affect mainly adjectives; events can be both nominal, verbal and adjectival. If we however extend this analysis to a language pair where negation is expressed through different means (e.g. English and Czech), it is unlikely we will find the same error distribution. Moreover, hierarchical phrase-based models are in fact non-purely syntax driven methods that are able to deal with high levels of reordering. That however also means that (a) there is no concept of syntactic constituent boundaries and (b) when reordering is performed incorrectly there is a high degree of element scrambling. For this reason phrase-based systems (where reordering is limited) and syntax-based systems (where an explicit knowledge of constituent boundaries is present) are likely to yield different results.

Finally, this paper has only discussed manual detection of translation errors involving negation. Other ongoing work tries instead to automate this process.

9 Acknowledgements

This work was supported by the Accept, MosesCore and GRAM+ grant. The authors would like to thank Adam Lopez for his comments and suggestions and the three anonymous reviewers for their feedback.

References

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

Blanco, Eduardo and Moldovan, Dan (2011). Some Issues on Detecting Negation from Text. In *Proceedings of the 24th Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*, pages 228–233, Palm Beach, FL, USA.

Chiang, David (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Chowdhury, Md and Mahbub, Faisal (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.

Collins, Michael and Koehn, Philipp and Kučerová, Ivona (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 531–540.

Fancellu, Federico and Webber, Bonnie (2012). 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 Webber, Bonnie (2014). Applying the semantics of negation to SMT through n-best list re-ranking. *EACL 2014*, page 598.

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

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

Hardmeier, Christian and Tiedemann, Jörg and Nivre, Joakim (2014). Translating pronouns with latent anaphora resolution. In *NIPS 2014 Workshop on Modern Machine Learning and Natural Language Processing*.

Li, Jin-Ji and Kim, Jungi and Kim, Dong-Il and Lee, Jong-Hyeok (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.

Lo, Chi-kiu and Wu, Dekai (2010). Evaluating machine translation utility via semantic role labels. In *Seventh International Conference on Language Resources and Evaluation (LREC-2010)*, pages 2873–2877.

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

Morante, Roser and Blanco, Eduardo (2012). *SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 265–274.

Morante, Roser and Schrauwen, Sarah and Daelemans, Walter (2011). Annotation of negation cues and their scope, Guidelines v1.0. *Computational linguistics and psycholinguistics technical report series, CTRS-003*.

Och, Franz Josef (2003). Minimum error rate training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167.

Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Read, Jonathon and Velldal, Erik and Øvrelid, Lilja and Oepen, Stephan (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.

Snover, Matthew G and Madnani, Nitin and Dorr, Bonnie and Schwartz, Richard (2009). TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3):117–127.

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

Vilar, David and Xu, Jia and Haro, Luis Fernando and Ney, Hermann (2006). Error analysis of statistical machine translation output. In *Proceedings of LREC*, pages 697–702.

Wetzel, Dominikus and Bond, Francis (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.

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