Enriching Parallel Corpora for Statistical Machine Translation with Semantic Negation Rephrasing

Dominikus Wetzel
Department of Computational Linguistics
Saarland University
dwetzel@coli.uni-sb.de

Francis Bond
Linguistics and Multilingual Studies
Nanyang Technological University
bond@ieee.org

Abstract

This paper presents an approach to improving performance of statistical machine translation by automatically creating new training data for difficult to translate phenomena. In particular this contribution is targeted towards tackling the poor performance of a state-of-the-art system on negated sentences. The corpus expansion is achieved by high quality rephrasing of existing sentences to their negated counterparts making use of semantic transfer. The method is designed to work on both sides of the parallel corpus while preserving the alignment. Our results show an overall improvement of 0.16 BLEU points, with a statistically significant increase of 1.63 BLEU points when tested on only negated test data.

1 Introduction

Having large and good quality parallel corpora is vital for the quality of statistical machine translation (SMT) systems. However, these corpora are expensive to create. Furthermore, certain phenomena are not very frequent and hence underrepresented in existing parallel corpora, such as negated sentences, questions, etc. Due to the lack of such training data, the SMT systems do not perform as well as they could. Especially when it comes to negation, it is important that the basic semantics is preserved, i.e. a negated statement should not be translated as a positive one and vice versa.

Given a state-of-the-art baseline Japanese-English SMT system, a separate evaluation on the semantic level of negative only vs. positive only test data reveals the considerably poorer performance on the negative test set. This tendency and the importance of preserving a negated statement motivates experiments with improving performance on negative sentences.

Providing more training data for negative sentences should even out the discrepancy of the performance between the above mentioned negative and positive test data. We present a method where a large amount of negative training data is obtained by rephrasing the original training data. The rephrasing is performed on the semantic level to ensure high reliability and quality of the generated data. Simple rewriting based on the surface or syntactic level would require complex language specific rules, which is not desirable.

Working on the semantic structure exploits the fact that these representations abstract away from language specific structures. Thus, our approach can be easily implemented for other languages, provided there are grammars available for both languages involved in the desired parallel corpus. The DELPH-IN project[1] provides various such grammars.

This paper first describes related work in the following section. Section 3 presents a semantic analysis of the data with respect to negation and provides some distributional statistics. In Section 4 we elaborate on the functionality of our rephrasing system and present different methods for corpus expansion. The experimental setup and the results are in Section 5. A discussion and our conclusion are given in Section 6 and Section 7, respectively.

2 Related Work

There has been plenty of work on paraphrasing data in order to overcome the limitations that insufficiently large or underrepresented phenomena in par-

[1] www.delph-in.net

Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 20–29,
Jeju, Republic of Korea, 12 July 2012.

allel corpora impose on SMT.

Callison-Burch et al. (2006) tackle the problem of unseen phrases in SMT by adding source language paraphrases to the phrase table with appropriate probabilities. Both are obtained from additional parallel corpora, where the translations of the same foreign language phrase are considered paraphrases.

He et al. (2011) use a statistical framework for paraphrase generation of the source language. A log-linear model similar to the one used in phrase-based SMT provides paraphrases which are ranked based on novelty and fluency. The training corpus is then expanded by either adding the first best paraphrase, or n-best paraphrases. The target language is just copied to provide the required target side of the paraphrase.

Marton et al. (2009) and Gao and Vogel (2011) create new information by means of shallow semantic methods. The former present an approach to overcome the problem of unknown words in a low resource experiment. They base their monolingual paraphrasing on semantic similarity measures. In their setting they achieve significantly better translations. Gao and Vogel (2011) expand the parallel corpus by creating new information from existing data. With the use of a monolingual semantic role labeller one side of the parallel corpus is labelled. Role-to-word rules are extracted. In sentences containing the frames and semantic roles for which replacement rules exist, the corresponding words are substituted. A support vector machine is used for filtering the generated paraphrases.

An approach where paraphrases are obtained via generation from semantic structures is presented in Nichols et al. (2010). It exploits the fact that the generator produces multiple surface realizations. The basic set up is similar to our work, however our approach additionally manipulates, i.e. rephrases the semantics before generation. Furthermore, we implement parallel rephrasing, changing the meaning of both source and target text simultaneously.

There is, on the other hand, little work in phrase-based SMT especially targeting negated sentences. Collins et al. (2005) approach the problem of properly translating negation in their general reordering setting. Transformation rules are applied to syntactic trees, so that the source language word order has a closer resemblance to the target language word order. In particular, the German negation is moved towards the same position as the English one. This however presumes the existence of at least some negated training data.

3 Analysis of the Semantic Structure

The linguistic analysis is performed based on the Head-Driven Phrase Structure Grammar (HPSG) formalism established in the DELPH-IN project. In particular we consider the language pair Japanese-English. Hence, the broad-coverage grammar Jacy for Japanese (Bender and Siegel, 2004) and the English Resource Grammar (ERG) (Flickinger, 2000) are used respectively to parse the data and obtain the semantics for each sentence.

3.1 Negation in Minimal Recursion Semantics

The formalism that is used to represent the semantics in the DELPH-IN grammars is Minimal Recursion Semantics (MRS) (Copestake et al., 2005). Per definition, an MRS structure consists of a top handle, a bag of elementary predicates (EP) and a bag of constraints on handles. EPs represent verbs, their arguments, negations, quantifiers, among others. Furthermore, each EP has a handle with which it can be identified. Constraints on handles are used to restrict EPs such that they are outscoped by negations or quantifiers.

In a negated sentence, the negated verb is outscoped by the negation relation EP. Technically, the negation relation with handle h_n takes as its argument (ARG1) a handle (h_x) which is equal modulo quantifiers to the handle of the verb (h_v), written as the handle constraint: $h_x =_q h_v$. For visualization, an example is given, which shows the relevant parts of such a negated structure for the sentence *"This may not suit your taste."* (Figure 1). There, the negated verb has the handle h_8. The negation relation EP with handle h_{10} outscopes this via the constraint $h_{12} =_q h_8$.

The rephrasing we propose can be achieved with little or no knowledge about the specific implementation choices of the individual grammar. Collecting a few sample sentences that appear to be negated in the original data – by performing a simple surface string matching – is enough to reveal the principle of how negation is implemented. Because negation

```
< e2,
  { h8: _MAY_V_MODAL_REL( ARG0 e2, ARG1 h9 ),
    h10: NEG_REL( ARG0 e11, ARG1 h12),
    h13: _suit_v_1_rel( ARG0 e14, ARG1 x4, ARG2 x15),
    ... }
  { h6 =q h3,
    h12 =q h8,
    h9 =q h13,
    ... } >
```

Figure 1: A visualization of the English MRS structure from the sentence *"This may not suit your taste."*. The irrelevant parts have been omitted. The necessary parts in the corresponding Japanese MRS are the same.

	Japanese	
English	neg_rel	no neg_rel
neg_rel	8.5%	1.4%
no neg_rel	9.7%	80.4%

Table 1: Distribution of negation measured by the presence or absence of a negation relation (*neg_rel*) for those sentences with parses in both languages.

is represented at the semantic level, both the ERG and Jacy have very similar analyses, even though the syntactic realization is very different (negation in English involves a negative marker such as *not* and the use of an auxiliary verb such as *do*, while in Japanese it is realized by an auxiliary verb *nai*).

3.2 Data and Distribution of Negations

The data we use in this work is the Japanese-English parallel Tanaka corpus (Tanaka, 2001; Bond et al., 2008). We used the version distributed with Jacy, which has approximately 150,000 sentence pairs randomly ordered and divided into 100 profiles of 1,500 sentences each (the last one is a little short). We summarize the distribution of negated sentence pairs in Table 1. The data we consider for these statistics excludes development and test profiles (000–005). 84.5% of the input sentence pairs can be parsed successfully (110,759 out of 139,150).

The table also shows mixed cases where one language had a negation relation EP, whereas the other did not. Mixed cases are especially frequent when the Japanese side has a negation relation. These cases have two main causes: lexical negation such as "She missed the bus." being translated with the equivalent of "She did not catch the bus."; and idioms, such as *ikanakereba naranai* "I must go (lit: go-not-if not-become)" where the Japanese expression of modality includes a negation. Instances of the latter type form the majority, and should be handled in a newer version of the grammar, they are not considered further in this work.

4 Method: MRS Rephrasing & Corpus Expansion

The basic setup of the whole rephrasing system consists of parsing, MRS manipulation, generation and finally parallel corpus compilation. In the following sections, the individual processing modules are described in detail.

4.1 Parsing

Parsing is done using PET (Callmeier, 2000) a bottom-up chart parser for unification-based grammars using the English and Japanese Grammars ERG and Jacy. Since our approach builds on semantic rephrasing, only the MRS structure is required. We only use the best (first) parse returned by the parser.

4.2 Rephrasing

This module takes an MRS structure as input and rephrases it if possible by adding a negation relation EP to the highest scoping predicate. Adding the negation relation in our current form does not explore alternatives, where the negation has scope over

other EPs in the MRS, nor are more refined changes from positive to negative polarity items considered.

Before inserting the negation relation EP into the existing MRS structure with its required handle constraint, we have to identify the EP we want to negate. The event that is introduced by the highest scoping verb is used. The event variable e_2 is directly accessible at the top of the MRS structure (cf. Figure 1). The corresponding EP that we want to negate has the event variable as value of its ARG0 attribute. This EP has a handle h_8 that has to be outscoped by the negation by means of a handle constraint. Hence, a new negation relation EP (in the example it got the handle h_{10}) is inserted with the following condition: Its ARG1 attribute value has to be token identical to the left side of a $=_q$ constraint. The right side is set to the just identified handle h_8 of the verb.

4.3 Generation

The same grammars used for parsing can also be used by the generator of the Lexical Knowledge Builder Environment (Copestake, 2002) to generate an n-best list of surface realizations given an MRS structure. However, we only consider the highest ranked realization. For the English generation, a generation ranking model is provided within the DELPH-IN project, thus providing a more confident n-best list. For the current Japanese grammar, no such model is available.

An example of a successful generation can be found in Table 2. On the English side, two surface variations are generated. The Japanese realizations show more variations in honorification and aspect.

We can only negate sentence pairs in both languages for 13.3% of the training data (18,727). This is mainly because of the brittleness of the Japanese generation (Goodman and Bond, 2009). Further, there are multiple ways of negating sentences and we do not always select the correct one.

4.4 Expanded Parallel Corpus Compilation

The method for assembling the expanded version of the parallel corpus for the use as training or development data directly influences translation quality. This is also demonstrated in Nichols et al. (2010), where various versions of padding out the data and preserving the word distribution are compared. The reported differences in performance suggest the importance of the method. Therefore, we have experimented with the following versions:

- **Append**: The obtained negated sentence pairs are added to the original corpus. Only the highest ranked realization per sentence for each language is considered. Thus they are aligned with each other. This leads to the addition of the following sentence pair where bilingual negation was successful:
  `(en_original,jp_original)`
  `(en_negated_1,jp_negated_1)` added
- **Padding**: In order to preserve the word distribution as mentioned above, we additionally padded out the sentence pairs by copying, where no bilingual negation was possible:
  `(en_original,jp_original)`
  `(en_original,jp_original)` added
- **Replace**: For emphasizing the impact of negated sentences, a variant of *Append* was compiled. Instead of adding the original pair of a successful bilingual negation the former was replaced by the latter:
  `(en_negated_1,jp_negated_1)` substituted

Another way of testing the quality of the generated rephrases is to include them in the language model training. The expectation is that when the rephrases are of good quality, then the language model will be better and in turn should have positive result on the overall SMT.

5 Experiments & Evaluation

We experiment with the phrase-based statistical machine translation toolkit Moses (Koehn et al., 2007) in order to train a Japanese - English system and to show the influence of the expanded parallel corpora obtained with negation rephrasing on translation performance.

5.1 Data

The Tanaka corpus is used as a basis for our experiments. We tokenize and truecase the English side, the Japanese side is already tokenized and there are no case distinctions. Sentences longer than 40 tokens are removed. For evaluation, the English part is recased and detokenized.

	English	Japanese
original	I aim to be a writer.	私 は 作家 を 目指し て いる 。
negated	I don't aim to be a writer.	私 は 作家 を 目指し て い ない
	I do not aim to be a writer.	私 は 作家 を 目指し て い ませ ん
		私 は 作家 を 目指し ませ ん
		私 は 作家 を 目指さ ない
		作 家 を 私 は 目指し ませ ん
		作 家 を 私 は 目指さ ない

Table 2: English and Japanese generations of a successfully rephrased sentence pair.

The sentence and token statistics for the original Tanaka corpus and our various extensions are listed in Table 3. The original corpus version acts as baseline data with profiles 006–100 as training and 000–002 as development data. For the extended systems, the training data as described in Section 4.4 is used. The same methods are applied on the development portion of the Tanaka corpus for tuning. The full test data has 42,305 English and 53,242 Japanese tokens and 4,500 sentences and is equal to the Tanaka corpus profiles 003–005.

The language model training data is in almost all cases equal to the original English Tanaka training data. Only in the *Append + neg LM* experiment, the training data for the language model is equal to the *Append* training data, except that it is slightly larger, since long sentences have not been filtered out. The expanded language model training data consists of 1,476,231 tokens and 160,069 sentences.

5.2 Different Test Sets

In order to find out the performance of the baseline and the extended systems on negative sentences, the test data has to be split up into several subsets, most notably *neg-strict* and *pos-strict*. The former only contains negated sentences, the latter only positive sentences. The definition of both is based on the existence of a negation relation EP in the semantics of the sentence. In order to obtain the semantic structure, the sentence pairs have to be parsed successfully. This also means, we will have some sentence pairs for which we cannot make a decision. Therefore, we provide a third test subset *biparse*, which contains all the parsable sentence pairs. This set reveals the big jump of BLEU score compared to the fourth test set *all*, which is the regular test set of the Tanaka corpus. A combined dataset with *pos-strict-neg-strict* is provided, which is the union of the first two sets.

5.3 Setup

We use Moses (SVN revision 4293) with Giza++ (Och and Ney, 2003) and the SRILM toolkit 1.5.12 (Stolcke, 2002). The language model is trained as a 5-order model with Kneser-Ney discounting. The Giza++ alignment heuristic *grow-diag-final-and* is used. All systems are tuned with MERT (Och, 2003). Several tunings for each system are run, the best performing ones are reported here.

5.4 Results

The results of our experiments can be seen in Table 4. The baseline is outperformed by our two best variations *Append* and *Append + neg LM* with respect to the entire test set. The differences in BLEU points are 0.14 and 0.16, which are not statistically significant according to the paired bootstrap resampling method (Koehn, 2004).

When looking at the test set *neg-strict* that only contains negated sentences, our improvement is much more apparent. The gain of our best performing model *Append + neg LM* compared to the baseline is at 1.63 BLEU points, which is statistically significant ($p < 0.05$). On the other hand there is a statistically insignificant drop of 0.30 with *pos-strict*.

The model with the expanded language model training data (*Append + neg LM*) always performs

	Tokens		Sentences	
	train	dev	train	dev
Baseline	1,300,821 / 1,641,591	42,248 / 52,822	141,147	4,500
Append	1,469,569 / 1,841,139	47,905 / 59,400	159,874	5,121
Padding	2,628,757 / 3,293,246	85,422 / 105,952	282,294	9,000
Replace	1,327,936 / 1,651,655	43,174 / 53,130	141,147	4,500

Table 3: Counts of tokens and sentences of the original Tanaka corpus and our expanded versions. Tokens are split up in English/Japanese counts.

better than the model under the same conditions except language model training data (*Append*).

When padding out the original data to preserve the word distribution in *Padding*, the effect of the additional negated training pairs is not strong enough. Both scores on the entire test set, as well as on the negation specific test set drop below the baseline. This version performs slightly better overall compared to *Replace*, however, on *neg-strict* it is a lot worse.

We manually checked the *neg-strict* test data set of our best performing system *Append + neg LM* versus the baseline, checking only whether the negation was translated or not (ignoring the overall quality). For 146 sentences, both systems correctly translated the negation. For 76 sentences both systems failed to translate the negation. For 33 sentences *Append + neg LM* translated the negation where the baseline system did not, and for 30 sentences the baseline system translated the negation but *Append + neg LM* did not. Overall, we reduced the number of critical negation errors from 99 to 96. Some example sentences are given in Figure 2.

6 Discussion

For identifying the performance of a state-of-the-art baseline system on negated sentences, we have split the test data into several distinct sets. The translation quality drops considerably by about 3 BLEU points when looking at the negative data compared to the parsable test data *biparse*. This big decline and the difference between performance on negative vs. positive test data shows that there is great potential to improve SMT systems by tackling this problem. Our approach is successful in handling negations better and thus diminishing the discrepancy of the two sets.

As the results show, there is only a small decrease of BLEU score points on the positive test data. And on the negative test data, the increase is substantially higher. Nevertheless, the overall performance in terms of BLEU only reflects this high increase to a certain degree. This can be attributed to the fact that the test data has a similar distribution to that of the training data, i.e. the proportion of negative sentences is low. Thus, the big increase gets diluted in the overall test data.

The results further show that improvement on the negative test data set comes at the cost of a slight degradation of performance on the positive data set and hence also on the full test set. This behaviour is not surprising due to the fact that a positive and its negative correspondent only vary very little when looking at the surface structure. The models trained with our extended data are aimed at providing one model which provides a balance between this gain and the loss.

This notion suggests that one would benefit from providing two separate translation models, one for negated input data and one for positive data. In this setting, the ample amount of negative training data that we generated through rephrasing could be exploited even more. A yet higher increase of BLEU score is expected. This of course requires a preprocessing step that confidently splits up the data accordingly. However, since we have the grammars at hand that can reliably determine whether there is a semantic negation relation in the input, this step can be solved easily. One small disadvantage with this idea is that a decision can only be made if the gram-

Test data sets	all	biparse	neg-strict	pos-strict	pos-strict-neg-strict
Sentence counts	4500	3399	285	2684	2969
Baseline	22.87	25.76	22.77	**26.60**	26.25
Append	23.01	25.78	24.04	26.22	26.25
Append + neg LM	**23.03**	**25.88**	**24.40**	26.30	**26.28**
Padding	22.74	25.54	22.62	26.35	26.06
Replace	22.55	25.35	23.36	26.00	25.84

Table 4: Japanese-English translation evaluation results of the baseline and our extended systems.

mar of the input language produces a parse for the input sentence. This however can be circumvented by backing off to the well balanced model presented in this work. In other words, we use a positive model for positive sentences, a negative model for negative sentences and a balanced model if we are not sure.

Our method depends on two large-scale deep semantic grammars. However, developing such grammars has been made much more efficient with the emergence of the Grammar Matrix (Bender et al., 2002). There is is already a large collection of working grammars, which can readily be tried out. In addition to the ERG and Jacy, there are grammars for German, French, Korean, Modern Greek, Norwegian, Spanish, Portuguese, and more, with varying levels of coverage.[2]

Because parsing, rephrasing and generation do not have 100% coverage, we cannot produce negated versions of all sentences. The rephrasing can only work when both sides of a sentence pair are parsable. Furthermore, not every rephrased sentence pair can be successfully realized. However, we still manage to build far more negated training data than is otherwise available: more than doubling the amount. This could be further increased by a little more work on the generation, especially for Jacy. In addition, we have not made use of all the generated data, i.e. lower ranked realizations have been discarded even though they may still be useful.

Furthermore, we have shown in the experiment results that using our expanded version for language model training is also of great benefit, since we could achieve not only an overall increase, but especially one on negated test data.

[2] `moin.delph-in.net/GrammarCatalogue`

7 Conclusion & Future Work

We have presented an approach which alleviates the negation translation difficulties of phrase-based SMT. We have tackled the problem by automatically expanding the training data with negated sentence pairs. The additional data has been obtained by rephrasing existing data based on the semantic structure of the input.

Our experiments with the phrase-based SMT system Moses show small improvements over the baseline considering the entire test data. A more distinct look at only negated sentences in the test data shows a statistically significant improvement of 1.63 BLEU points. The best performing model represents a good balance of a high BLEU score increase on the negated test data vs. a statistically insignificant decrease on the positive test data, yet achieving a small overall improvement. Furthermore, it was shown, that expanding not only the translation training data, but also the language model training data boosts performance even more.

Our method works on the semantic level and can be easily adapted to other languages. Having access to a deep semantic structure opens possible extensions along our idea. On the one hand negation rephrasing could be refined in order to have a higher generation rate. On the other hand, other phenomena could also be tackled in the same way: e.g. rephrasing declarative statements to interrogatives.

Just for negation, the corpora expanded with our high quality negations could be combined with the syntactic reordering strategies presented in Section 2 such that the negation reordering rule has more training data and thus a bigger influence on the overall performance.

Acknowledgements

This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2009-5259-5.

References

Bender, E. M., Flickinger, D., and Oepen, S. (2002). The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

Bender, E. M. and Siegel, M. (2004). Implementing the syntax of Japanese numeral classifiers. In *Proceedings of the IJC-NLP-2004*.

Bond, F., Kuribayashi, T., and Hashimoto, C. (2008). Construction of a free Japanese treebank based on HPSG. In *14th Annual Meeting of the Association for Natural Language Processing*, pages 241–244, Tokyo. (in Japanese).

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics.

Callmeier, U. (2000). PET - a platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering*, 6(1):99–108.

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, Ann Arbor, Michigan. ACL.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications.

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

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28. (Special Issue on Efficient Processing with HPSG).

Gao, Q. and Vogel, S. (2011). Corpus expansion for statistical machine translation with semantic role label substitution rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 294–298, Portland, Oregon, USA. Association for Computational Linguistics.

Goodman, M. W. and Bond, F. (2009). Using generation for grammar analysis and error detection. In *Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 109–112, Singapore.

He, W., Zhao, S., Wang, H., and Liu, T. (2011). Enriching smt training data via paraphrasing. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 803–810, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the ACL*.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Singapore. Association for Computational Linguistics.

Nichols, E., Bond, F., Appling, D. S., and Matsumoto, Y. (2010). Paraphrasing Training Data for Statistical Machine Translation. *Journal of Natural Language Processing*, 17(3):101–122.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of ACL*, pages 160–167.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51.

Stolcke, A. (2002). SRILM – An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904, Denver.

Tanaka, Y. (2001). Compilation of a multilingual parallel corpus. In *Proceedings of PACLING 2001*, pages 265–268, Kyushu.

Japanese	昨日 彼ら は テニス を し なかっ た 。
Baseline	They played tennis yesterday.
Append + neg LM	They do not play tennis yesterday.
Reference	Yesterday they didn't play tennis, because it rained.

(a) Baseline fails to translate the negation.

Japanese	彼 は 約束 を 破る こと は し ない と 確信 し て いる ん です が 。
Baseline	He is sure to break your promise, I'm sure.
Append + neg LM	He never breaks his word, I'm sure.
Reference	I'm sure he won't fail to keep his word.

(b) Correct translation by our system with valid variation of wording.

Japanese	私 が 家 に 帰っ た 時 は 彼 は 眠っ て い ませ ん でし た 。
Baseline	I was when I came home, he was asleep.
Append + neg LM	I came home when he is not asleep.
Reference	He wasn't sleeping when I came home.

(c) Baseline omits the negation.

Japanese	お金 の もちあわせ が あり ませ ん 。
Baseline	Money with me.
Append + neg LM	I don't have any money with me.
Reference	I don't have any money with me.

(d) Baseline omits subject, verb and negation.

Japanese	南十字星 は 日本 で は 見る こと が でき ない 。
Baseline	The 南十字星 in Japan, I cannot see it.
Append + neg LM	The 南十字星 in Japan.
Reference	The Southern Cross is not to be seen in Japan.

(e) Our system does not translate a part of the sentence.

Japanese	大声 で 話し て は いけ ない 。
Baseline	Don't speak in a loud voice.
Append + neg LM	You must speak in a loud voice.
Reference	You must not speak loudly.

(f) Our system omits the negation.

Japanese	彼女 は 友達 が い ない 。
Baseline	She has no friends.
Append + neg LM	She is a friend of mine.
Reference	She doesn't have a boy friend.

(g) Our system does not produce a negation. The object is incorrectly translated in both systems.

Figure 2: Sentences from the *neg-strict* test set showing differences between the baseline and our best performing system *Append + neg LM*. Examples in (a–d) show improvements, (e–g) show degradations.