Promoting Flexible Translations in Statistical Machine Translation

Rico Sennrich
Institute of Computational Linguistics
University of Zurich
Binzmühlestr. 14
CH-8050 Zürich
sennrich@cl.uzh.ch

Abstract

While SMT systems can learn to translate multiword expressions (MWEs) from parallel text, they typically have no notion of non-compositionality, and thus overgeneralise translations that are only used in certain contexts. This paper describes a novel approach to measure the flexibility of a phrase pair, i.e. its tendency to occur in many contexts, in contrast to phrase pairs that are only valid in one or a few fixed expressions. The measure learns from the parallel training text, is simple to implement and language independent. We argue that flexible phrase pairs should be preferred over inflexible ones, and present experiments with phrase-based and hierarchical translation models in which we observe performance gains of up to 0.9 BLEU points.

1 Introduction

A defining property of multiword expressions (MWEs) is that they are idiosyncratic (Sag et al., 2002). For Statistical Machine Translation (SMT), MWEs whose meaning is non-compositional, i.e. which cannot be translated word by word, can cause two major problems. The obvious problem is that MWEs may be translated incorrectly if we translate them word by word. A second problem, which has received less attention in SMT research, is that translations that we learn from the components of a MWE can rarely be generalised to other contexts. If a word frequently occurs in a MWE with an idiosyncratic translation, learning this idiosyncratic translation on the word level pollutes the translation model.

Consider for instance the English phrase *of course*, which is translated into French as *bien sûr*. SMT systems, which typically perform unsupervised word alignment to learn translation correspondences, not only learn the translation pair (*of course*, *bien sûr*), but also (*of*, *bien*) and (*course*, *sûr*). Especially if the fixed expression *of course* is more frequent during training than other translations of *course*, the translation pair (*course*, *sûr*) is misapplied to occurrences of *course* in new contexts. This problem affects various linguistic phenomena that fall under the umbrella term MWE: complex prepositions, idioms, compounds and named entities, among others.

We describe an algorithm to measure a phrase pair's flexibility, i.e. whether it occurs in many contexts or is restricted to fixed expressions. Note that the aim is not to penalize MWEs themselves, which may be flexible in terms of their contexts, but only phrases that are part of a larger MWE. In contrast to other related work on MWEs in SMT, our approach is unsupervised and language independent.

2 Related Work

The fact that word-based translation techniques are inadequate to deal with MWEs, which are by definition non-compositional, has led to approaches that extract MWEs in order to improve bilingual resources (e.g. (Smadja et al., 1996; Carpuat and Diab, 2010)). Using contextual information to disambiguate translations is an equally well-researched topic (Carpuat and Wu, 2007; Chiang et al., 2009). One can even argue that the success of phrase-based SMT (Koehn et al., 2003) compared

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 207–214.

to word-based approaches is in large part due to the existence of MWEs in natural language.

Our work is not concerned with improving the translation of MWEs themselves, but with preventing an overgeneralisation of translations learned from MWEs. In other words, our aim is not to improve the translation of words in known contexts, but picking a better translation if a word occurs in a new context. It shares the aim with the work by Lambert and Banchs (2006), who convert MWEs into single tokens in a preprocessing step, thus preventing MWE sub-segments from being extracted. In order to identify MWEs in the source text, they exploit asymmetries in word alignment, lemmatisation and PoS-tagging. They found that the positive effect of suppressing wrong phrase pairs in some instances was counterbalanced by increased data sparseness, especially because some word sequences were erroneously identified as MWEs. Pal et al. (2011) follow the same idea for the language pair English–Bengali, with different MWE extraction techniques.

3 Learning Translations in SMT

To illustrate why wrong translations are learned from MWEs, let us consider the common SMT training process. In (hierarchical) phrase-based SMT, translations are extracted from a word-aligned corpus. This extraction is performed by heuristics that extract phrase pairs which are consistent with word alignment, specifically, so that no word in the source phrase is aligned to a word outside the target phrase, and vice versa. For MWEs, this means that phrase pair extraction for the whole MWE, and its subphrases and words, are co-dependent. A MWE is only extracted if its components do not violate word alignment, and when the latter is the case, this also entails that these components will form phrase pairs of their own. In other words, the phrase table is learned with a compositionality assumption, and the model has no means to learn that a phrase pair can be correct while its components should not be used independently.

While state-of-the-art SMT systems have this technical weakness, they are easy to extend thanks to their log-linear framework. In the final translation model, each extracted phrase pair $(\overline{s}, \overline{t})$ has multiple scores, which are combined with each other and other features such as the language model probability in a log-linear model. Most common are phrase translation probabilities estimated through (smoothed) relative frequencies $p(\overline{s}|\overline{t})$ and $p(\overline{t}|\overline{s})$, and a smoothed probability distribution based on word translation probabilities (Koehn et al., 2003). We extend this log-linear model through new features that measure a phrase pair's flexibility.

4 Flexibility Features

We introduce new probability distributions that are not based on relative frequency estimates, but on the number of different contexts in which a phrase pair occurs. Intuitively, we use them to predict how likely a phrase pair is to occur in a new context. We will call a phrase pair flexible if it occurs in many contexts, as opposed to inflexible phrase pairs that we only observe in few contexts. Note that under this definition, even fixed expressions may be considered flexible if they themselves occur in many contexts. It is not the translation of MWEs that we aim to penalize, but the translation of their individual segments.

In order to measure a phrase pair's flexibility, we introduce equation 1. Given a source phrase $\overline{s}$ and a target phrase $\overline{t}$, with $\overline{s}$ being a sequence of words from s_i to s_j, we consider triplets of the form $(s_x, \overline{s}, \overline{t})$ for the flexibility measure. Different positions can be considered for s_x. We introduce two new probability distributions; the first, with $x = i - 1$, is based on the number of contexts to the left of $\overline{s}$, and will be referred to as $p_{\text{flex\_left}}$. The second, $p_{\text{flex\_right}}$, is based on the number of right contexts, with $x = j + 1$.[1]

$$\begin{aligned} p_{\text{flex\_\{left,right\}}}(\overline{t}|\overline{s}) &= \frac{N_{1+}(\bullet, \overline{s}, \overline{t})}{\sum_{\overline{t}'} N_{1+}(\bullet, \overline{s}, \overline{t}')} \\ &= \frac{N_{1+}(\bullet, \overline{s}, \overline{t})}{N_{1+}(\bullet, \overline{s}, \bullet)} \end{aligned} \quad (1)$$

N_{1+} denotes the number of types, and $\bullet$ are wildcards. $N_{1+}(\bullet, \overline{s}, \overline{t})$ is thus the number of different triplets $(s_x, \overline{s}, \overline{t})$ observed during training. $p_{\text{flex\_left}}(\overline{s}|\overline{t})$ and $p_{\text{flex\_right}}(\overline{s}|\overline{t})$ are calculated analogously, i.e. by considering the number of contexts to the left and right of the target phrase. We can theoretically increase the window that we consider to be the context of a phrase, but as we increase

[1] We add a special token for s_{i-1} if the phrase begins at the start of a sentence, and do the same for s_{j+1} at the end.

the window size, the number of different types increases, and the new probability estimates tend towards the baseline relative frequency estimate.

Table 1 shows the effect of this calculation for selected phrase pairs.[2] The first example illustrates how the English complex preposition *in line with* is affected. In German, it is typically translated to *im Einklang mit*. However, if *line* occurs in other contexts, a typical translation would be *Linie* or *Reihe*, but almost never *Einklang*. The latter translation is restricted to *in line with*. We see that our model risks to translate *line* as *Einklang* because the relative-frequency estimate for $(\textit{Einklang}|\textit{line})$ is higher than that of $(\textit{Linie}|\textit{line})$. However, since the phrase pair (*line*, *Einklang*) occurs in very few contexts, both on the source and target side, its flexibility estimates are much lower than the estimate obtained from relative frequencies. $(\textit{Linie}|\textit{line})$ and $(\textit{im Einklang mit}|\textit{in line with})$ both occur in various contexts, and their flexibility estimates remain relatively high. The latter is an important point of our technique: the translation of MWEs as a single unit, which is a desirable property of phrase-based models, is not penalized.

The second set of phrase pairs are based on our introductory example (*of course*, *bien sûr*), and demonstrate why we measure flexibility to the right and to the left independently. The phrase pair $p(\textit{course}|\textit{sûr})$ is only inflexible to the left of *course*, but should still be penalized.

If a phrase pair is frequent, but only occurs in few contexts, this indicates that it is part of a larger MWE, and can safely be dispreferred. The cases in which an inflexible phrase pair is in fact a good translation should usually be handled by larger translation units, i.e. the MWE as a whole. However, there are exceptions to this rule. An exception are phrase pairs that typically occur at the beginning or end of a sentence. These may be inflexible according to our model, even if they are not part of a larger translation unit.

The flexibility measure has the same aim as the joining of MWEs that Lambert et al. (2006) describe, namely to prevent overgeneralisation of phrase pairs learned from MWEs to new contexts. It has the advantage of being language-independent and requiring no additional resources. Additionally, it does not need to make a hard classification into MWEs and others, and thus does not suffer from an increase in data sparseness.

4.1 Variants for Hierarchical Phrase-based Models

We can extend the notion of a phrase pair's flexibility to hierarchical phrase-based models (Chiang, 2005). However, we argue that a naive transfer of the approach to hierarchical phrase-based systems is incomplete. Because subphrases are allowed in hierarchical rules, there are additional ways in which a rule can be inflexible. Consider these three rules that might be learned from occurrences of the phrase pair (*of course*, *bien sûr*).

1. $X \rightarrow \langle$ *course* , *sûr* $\rangle$
2. $X \rightarrow \langle X_1$ *course* , X_1 *sûr* $\rangle$
3. $X \rightarrow \langle$ *and* X_1 *course* , *et* X_1 *sûr* $\rangle$

Each of these examples is a poor generalisation, and should ideally be penalized in the model. In all cases, we only expect the translation of *course* into *sûr* if *course* is preceded by *of*. In phrase-based models, this can be expressed through the relative number of left contexts observed with the source phrase, or $p_{\text{flex\_left}}(\bar{t}|\bar{s})$. This works for the hierarchical rule 1, but not for 2 and 3.[3] The reason is that *of* is not to the left of the rules extracted from the corpus, but part of the subphrase X_1. This leads to the question how we can formulate an alternative notion of context, so that the inflexibility of rules 2 and 3 can be learned.

We denote **FLEX_H1** the approach with $p_{\text{flex\_left}}$ and $p_{\text{flex\_right}}$ that has been described in the last section, and present multiple alternatives:

4.1.1 FLEX_H2

The first variant, called **FLEX_H2** henceforth, redefines which word is considered the left and right context in a hierarchical rule. If a hierarchical rule starts with a subphrase, the rightmost word of this subphrase is considered the rule's left context (instead of the word to the left of the subphrase). In other words, this variant does not use $x = i - 1$ for $p_{\text{flex\_left}}$, but $x = i - 1 + n$, with n being the length of the subphrase that the rule starts with, or $n = 0$ if the rule does not start with a subphrase. If it ends

[2] The examples are from the models described in section 6.1.

[3] In our training corpus from section 6.1, $p_{\text{flex\_left}}(\bar{t}|\bar{s})$ is $\frac{2}{2688}$ for rule 1, $\frac{480}{3506}$ for rule 2, and $\frac{232}{722}$ for rule 3. In other words, $p_{\text{flex\_left}}(\bar{t}|\bar{s})$ successfully assigns a low probability to the inflexible rule 1, but too high a probability to rules 2 and 3.

$\overline{s}$	$\overline{t}$	$p_{\text{RF}}(\overline{t}\|\overline{s})$	$p_{\text{flex\_left}}(\overline{t}\|\overline{s})$	$p_{\text{flex\_right}}(\overline{t}\|\overline{s})$
line	Einklang	0.159	0.003	0.005
line	Linie	0.146	0.072	0.061
in line with	im Einklang mit	0.169	0.073	0.059
course	sûr	0.444	0.001	0.066
course	cours	0.079	0.023	0.010

Table 1: Translation model probabilities for selected phrase pairs with different probability estimation methods: relative frequency (p_{RF}); flexibility distribution ($p_{\text{flex\_left}}$ and $p_{\text{flex\_right}}$).

with a subphrase, the leftmost word of this subphrase is considered for $p_{\text{flex\_right}}$, or $x = j + 1 - n$, with n being the length of the subphrase that the rule ends with, or $n = 0$ if the rule ends with a terminal symbol. This new definition aims to capture the inflexibility of rule 2.

4.1.2 FLEX_H3

A second possibility is to not only consider the words to the left and right of a rule to be its context, but also its subphrase(s). We start with **FLEX_H2**, and add a new feature $p_{\text{flex\_sub}}(\overline{t}|\overline{s})$ that is based on the number of different types of subphrases a hierarchical rule occurs with. This new feature is implemented through equation 1 by redefining s_x. Instead of the word to the left or the right of a subphrase, let s_x be defined as the full subphrase, or, if a rule contains multiple subphrases, the concatenation of all subphrases.[4] If a rule does not have a subphrase, we let $p_{\text{flex\_sub}}(\overline{t}|\overline{s})$ be 1, so that this feature is without cost for rules without subphrases. We also add $p_{\text{flex\_sub}}(\overline{s}|\overline{t})$, which is defined analogously.

5 Filtering Hierarchical Rule Tables

In preliminary experiments, we found that some of the differences between the baseline system and the experimental ones were due to spurious phrase or rule pairs whose probability estimates were unduly high. Thus, we use significance test filtering (Johnson et al., 2007) for phrase tables, which, as the authors note, has a similar effect as smoothing, since both pruning and smoothing penalizes infrequent phrase pairs. We extend their approach to hierarchical rule tables. Since (Johnson et al., 2007) do not base the significance test on alignment counts, but co-occurrence counts in the parallel corpus, we decided on an approximative method to count the number of occurrences of hierarchical rules, which can be implemented with a suffix array. Three frequencies are required to perform a statistical significance test for a phrase pair or rule: c_s, the frequency of the source phrase/rule, c_t, the target phrase/rule frequency, and c_{st}, the co-occurrence frequency of the source and target phrase/rule.

For hierarchical rules without subphrases, or which consist of a single, uninterrupted terminal sequence with subphrases at the beginning and/or end of the rule, we can use the same procedure as for phrase-based systems, namely extracting a set of sentences in which the source terminal sequence occurs, doing the same for the target sequence, and intersecting the two sets to obtain c_{st}.

For hierarchical rules which consist of multiple terminal sequences, interrupted by subphrases, we approximate its occurrences by extracting the set of occurrences for each terminal sequence, and using the intersection of these sets as occurrences of the full rule.

For a rule $X \rightarrow \langle\ a\ b\ X_1\ c\ , x\ X_1\ y\ z\ \rangle$, c_s is thus the number of source sentences in which *a b* and *c* occur, c_t the number of target sentences in which *x* and *y z* occur, and c_{st} the number of sentence pairs in which *a b* and *c* occur in the source sentence, *x* and *y z* in the target sentence.

6 Evaluation

6.1 Data and Methods

We perform the evaluation on the language pairs French–English and German–English, with training data mostly from the shared task of WMT 2011 (Callison-Burch et al., 2011). For both language pairs, we use Europarl and News-Commentary as parallel data sets. Language models are trained on the respective target language sides of Europarl, News-Commentary, and the monolingual News

[4] We insert a delimiter between two subphrases to distinguish between X_1 = 'a b', X_2 = 'c' and X_1 = 'a', X_2 = 'b c'.

system	DE–EN		EN–DE		FR–EN		EN–FR	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
newstest2011								
baseline	21.0	28.6	15.6	36.2	28.6	33.8	30.3	51.2
FLEX	**21.2**	**28.7**	**15.8**	**36.4**	**28.8**	33.8	30.3	51.1
cross-domain								
baseline	29.1	28.9	27.0	42.0	25.5	32.2	22.2	44.1
FLEX	**29.4**	**29.1**	27.1	**42.2**	**26.4**	**32.6**	**22.6**	**44.5**

Table 3: SMT results on newstest2011 and cross-domain test sets. Phrase-based models.

Data set	sentences	words (EN)
EN–FR		
News-commentary	110k	2900k
Europarl	1830k	50 600k
United Nations	11 800k	300 000k
10^9 corpus	21 400k	551 000k
EN–DE		
News-commentary	140k	3300k
Europarl	1740k	48 000k
JRC-Acquis	1200k	25 800k
OpenSubtitles v2	4650k	35 400k
News (EN)	110 000k	2 650 000k
News (FR)	25 000k	610 000k
News (DE)	52 000k	920 000k

Table 2: Training data used in evaluation.

data set, interpolated for minimal perplexity on $newstest2008$. For French–English, we additionally used the 10^9 corpus and United Nations corpus as parallel data sets. As additional German–English data, we used JRC-Acquis, a collection of legislative texts (Steinberger et al., 2006), and OpenSubtitles v2, a parallel corpus extracted from film subtitles[5] (Tiedemann, 2009). The respective data sizes are listed in table 2.

We train all systems with Moses (Koehn et al., 2007), SRILM (Stolcke, 2002), and GIZA++ (Och and Ney, 2003). We measure translation performance through BLEU (Papineni et al., 2002) and METEOR 1.3 (Denkowski and Lavie, 2011). All results are lowercased and tokenized, measured with five independent runs of MERT (Och and Ney, 2003) and using MultEval (Clark et al., 2011) to account for optimizer instability. Results marked in bold are statistically significantly better than the baseline according to significance testing in MultEval ($p < 0.05$).

We evaluate each system with two test sets. The first test set is newstest2011 from WMT 2011, with the system optimized on news-test2008; as second test set, we use patent abstracts for FR–EN[6], and help desk tickets provided to us by the software company Finnova for DE–EN. The reason for this is that we expect idiomaticity to be more of a problem if training and test set are dissimilar, since MWEs may be domain-specific (Smadja et al., 1996). We will refer to the second test set as *cross-domain* setting.

6.2 Phrase-based Results

Table 3 shows our experimental results with phrase-based systems on the newstest2011, and the two cross-domain test sets. The only change of our FLEX system over the baseline is the addition of four flexibility features to the log-linear model, namely $p_{\text{flex\_left}}$ and $p_{\text{flex\_right}}$ in both translation directions.

On newstest2011, we observe an improvement of 0.2 BLEU in three of the four translation directions. On the help desk and patent test sets, the flexibility features lead to larger improvements of up to 0.9 BLEU (FR–EN), with 0.3–0.4 points of improvement observed for DE–EN and EN–FR, and no significant improvement for the language pair EN–DE.

There are a number of possible explanations as to why we observe a gain in performance with some test sets, but not with others. Defining the context as the immediate neighbours of a phrase pair does have limitations. In the case of DE–EN, for instance, we note that the relatively free word order in German makes it harder to recognize if a word is part of a MWE with our approach. An

[5] `http://www.opensubtitles.org`

[6] extracted from the COPPA corpus (Pouliquen and Mazenc, 2011); IPC section A: human necessities.

system	sentence
source	Le jeu comprend des cartes objectifs et de l'argent pour le **jeu**.
reference	The game apparatus includes target cards, and **game** money.
baseline	The game includes maps objectives and money for the **stake**.
+FLEX	The game includes maps objectives and money for the **game**.
source	Improvements relating to **board games**
reference	Améliorations apportées à des **jeux de société**
baseline	Des améliorations relatives aux **jeux du conseil**
+FLEX	Des améliorations concernant les **jeux de société**

Table 4: Example translations from patents corpus. Phrase-based models.

example are German verb particles, which may occur not immediately after the verb, but at the end of the matrix clause. Without preordering, we cannot reliably distinguish between *schlägt* (engl: *beats*) and *schlägt ... vor* (engl: *proposes*).

Apart from the translation direction, the extent to which (parts of) MWEs are misused during translation depends on the training and test domain, since MWEs may be domain-specific (Smadja et al., 1996). If training and test domain are similar, using an idiomatic translation learned from the domain is more likely to be right than if the test set is from a different domain. Conversely, we expect cross-domain performance to benefit more strongly from the flexibility features, and consider this the main reason why we observe a larger performance boost with cross-domain test sets. Considering that we observe the largest performance gains in cross-domain translation, we argue that the flexibility features may be especially helpful for general purpose SMT system, and/or systems that use training data from various different domains.

Furthermore, adding flexibility features has side effects which may be both positive and negative. Specifically, the systems with flexibility features tend to give a higher weight to the phrase penalty in the log-linear model, meaning that the systems use fewer, but larger translation units during decoding. Such a preference of large translation units makes sense if we want to correctly translate MWEs despite the flexibility features: note our motivating examples in table 1, and that the flexibility features penalize (*Einklang*|*line*), but not (*im Einklang mit*|*in line with*).

Table 4 shows two examples where the baseline system misapplies an inflexible phrase pair. In these examples, the translations are so wrong that it might be hard to intuitively understand why they were even learned in the model. It is thus helpful to know the relevant phrase pairs, all frequent in the training set, that introduce these word translations into our model:

- *en **jeu** – at **stake***
- ***board** of directors – **conseil** de direction*

In both examples, the flexibility features successfully penalizes the (misused) idiomatic translation. However, the second example nicely illustrates that the experimental system does not prevent the translation of multiword expressions when they are encountered as a whole. The experimental system penalizes the inflexible translation pair (*conseil*|*board*), but not (*jeux de société*|*board games*), which is chosen instead. The example also illustrates our point about MWEs being domain-specific: *board of directors* only occurs once in a patent corpus of 9 million sentences, but 12 500 times in the 10^9 corpus (21 million sentences).

6.3 Hierarchical Results

Table 5 shows translation results for hierarchical systems. As far as statistical significance filtering is concerned, we see an increase in BLEU by up to 0.4 points for the filtered models, along with a reduction in rule table size. METEOR remains constant, or, for EN–DE, drops slightly by 0.2 points. A closer look at the METEOR statistics gives us an explanation for this discrepancy between BLEU and METEOR. Unigram precision benefits from significance filtering, while unigram recall, which is only considered by METEOR, is slightly decreased. We conduct all future experiments with filtered tables.

Just as with the phrase-based systems, the im-

system	DE–EN		EN–DE		FR–EN		EN–FR	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
newstest2011								
unfiltered	21.1	29.1	15.5	36.4	29.0	34.0	30.2	51.0
filtered	21.5	29.1	15.6	36.2	29.2	34.1	30.4	51.0
FLEX_H1	21.5	29.1	**15.8**	**36.5**	**29.6**	34.2	30.5	51.1
FLEX_H2	21.5	29.1	**15.9**	**36.5**	**29.5**	34.2	**30.6**	**51.2**
FLEX_H3	21.5	29.1	**15.8**	**36.5**	**29.7**	**34.3**	30.5	51.1
cross-domain								
filtered	29.2	29.1	26.3	41.6	24.2	31.9	22.8	44.7
FLEX_H1	29.3	29.2	**27.1**	**42.5**	**24.9**	**32.1**	22.7	44.7
FLEX_H2	29.3	29.2	**27.1**	**42.4**	**25.0**	**32.2**	22.8	44.7
FLEX_H3	29.3	29.2	**27.2**	**42.5**	**24.7**	32.0	22.8	44.8

Table 5: SMT results on newstest2011 and cross-domain test sets. Hierarchical models. Highlighted systems are significantly better than (filtered) baseline.

pact of the flexibility scores varies between the different translation directions and test sets. The biggest effect is observed for the translation directions EN–DE, with 0.2-0.3 BLEU points gained on newstest2011, and 0.8–0.9 BLEU points on the cross-domain test set, and FR–EN, with 0.3-0.5 BLEU points gained on newstest2011, and 0.5-0.8 BLEU points on the cross-domain test set. All variants of hierarchical flexibility scores perform similarly well, with no consistent winner variant. For DE–EN and EN–FR, adding flexibility scores yields no significant improvement.

A comparison between phrase-based and hierarchical systems gives a mixed picture. For the language pair FR–EN, the hierarchical system is better on newstest2011, the phrase-based one on the patent test set. An analysis of the METEOR statistics suggests that the highest difference between the phrase based and the hierachical models is in METEOR's fragmentation penalty, which means that reordering phenomena are at the root of these differences. Adding flexibility features is effective for both types of models; it primarily affects the precision and recall scores, which indicates that they improve the accuracy of the translation.

7 Conclusion

We describe a simple, yet effective way to measure the flexibility of phrase pairs, and show that these flexibility measures improve translation quality. By penalizing inflexible phrase pairs, i.e. phrase pairs that only occur in the context of larger multiword expressions, we measured gains in translation quality of up to 0.9 BLEU and METEOR points. The flexibility of phrase pairs is learned from the parallel training text, and expressed through new features in the log-linear SMT model. This makes the approach simple to implement and language-independent.

We have applied the approach to both phrase-based and hierarchical phrase-based SMT models, and discussed variants for hierarchical models. We have also demonstrated that rule table filtering based on statistical significance tests is possible and fruitful.

The flexibility scores that we proposed still have their limitations. Specifically, there are MWEs which have a higher flexibility on the surface level, but which we still would like to mark as inflexible. An example are German verb particles, which do not necessarily occur immediately after the verb, but at the end of the matrix clause, and whose translation is often non-compositional. Coupling flexibility scores with reordering might help to overcome these limitations.

Acknowledgements

I want to thank Martin Volk, Mark Fishel, and the anonymous reviewers for their valuable feedback. This research was funded by the Swiss National Science Foundation under grant 105215_126999.

References

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland.

Carpuat, Marine and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, CA.

Carpuat, Marine and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic.

Chiang, David, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 218–226, Boulder, CO.

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270, Ann Arbor, Michigan.

Clark, Jonathan H., Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 176–181, Portland, Oregon.

Denkowski, Michael and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland.

Johnson, Howard, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague, Czech Republic.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton, Canada.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Lambert, Patrik and Rafael Banchs. 2006. Grouping multi-word expressions according to part-of-speech in statistical machine translation. In *Proceedings of the EACL Workshop on Multi-Word-Expressions in a Multilingual Context*, pages 9–16, Trento, Italy.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pal, Santanu, Tanmoy Chakraborty, and Sivaji Bandyopadhyay. 2011. Handling multiword expressions in phrase-based statistical machine translation. In *Proceedings of the 13th Machine Translation Summit*, pages 215–224, Xiamen, China.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, Philadelphia, PA.

Pouliquen, Bruno and Christophe Mazenc. 2011. COPPA, CLIR and TAPTA: three tools to assist in overcoming the patent language barrier at WIPO. In *Proceedings of the 13th Machine Translation Summit*, pages 24–30, Xiamen, China.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 1–15, London, UK. Springer-Verlag.

Smadja, Frank, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1):1–38.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Stolcke, Andreas. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Seventh International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Tiedemann, Jörg. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia.