Polibits
ISSN: 1870-9044
polibits@nlp.cic.ipn.mx
Instituto Politécnico Nacional
México

Turchi, Marco; Ehrmann, Maud

Knowledge Expansion of a Statistical Machine Translation System using Morphological Resources

Polibits, vol. 43, 2011

Instituto Politécnico Nacional

Distrito Federal, México

Available in: http://www.redalyc.org/articulo.oa?id=402640456005

- How to cite
- Complete issue
- More information about this article
- Journal's homepage in redalyc.org

Scientific Information System
Network of Scientific Journals from Latin America, the Caribbean, Spain and Portugal
Non-profit academic project, developed under the open access initiative

# Knowledge Expansion of a Statistical Machine Translation System using Morphological Resources

Marco Turchi and Maud Ehrmann

***Abstract*—Translation capability of a Phrase-Based Statistical Machine Translation (PBSMT) system mostly depends on parallel data and phrases that are not present in the training data are not correctly translated. This paper describes a method that efficiently expands the existing knowledge of a PBSMT system without adding more parallel data but using external morphological resources. A set of new phrase associations is added to translation and reordering models; each of them corresponds to a morphological variation of the source/target/both phrases of an existing association. New associations are generated using a string similarity score based on morphosyntactic information. We tested our approach on En-Fr and Fr-En translations and results showed improvements of the performance in terms of automatic scores (BLEU and Meteor) and reduction of out-of-vocabulary (OOV) words. We believe that our knowledge expansion framework is generic and could be used to add different types of information to the model.**

***Index Terms*—Machine translation, knowledge, morphological resources.**

## I. INTRODUCTION

THE translation capability of a Statistical Machine Translation (SMT) system is driven by the training data and process. Big amounts of parallel data are used to allow the system to cover the source language as much as possible, but this effort collides with the vocabulary dimension of a language and the fact that the probability of finding unseen words in a language never vanishes. The inner knowledge of a system is the output of the training process that transforms the parallel data into tables: translation, language and reordering. Each item in translation and reordering tables associates textual (links phrase/s in different languages) and probability information (measures how reliable the information in the textual part is).

In real world translation systems, where source sentences may come from different domains, lack of knowledge is often responsible for translation quality: large number of OOV words or incorrect translations in target sentences are the main problems. In particular, when the source language is morphologically richer than the target language, translations are highly affected by the presence of OOV words. The other way around, the number of source phrases covered during the translation is higher, but target sentences contain more incorrect translated words.

Adding more data is the most obvious solution, but this has well-known drawbacks: it heavily increases the dimension of the tables, which reduces the translation speed, and parallel data are not always available for all the language pairs. In case of low quality parallel data, it can be even harmful because more data imply a bigger number of unreliable or incorrect associations built during the training phase.

In this paper, we address the problem of expanding the knowledge of an SMT system without adding parallel data, but extending the knowledge produced during the training phase. The main idea consists of inserting artificial entries in the phrase and reordering models using external morphological resources; the goal is to provide more translation options to the system during the construction of the target sentence.

Given an association of the phrase table, we first expand the source and target phrases, generating all their possible morphological variations. Then, given two sets of filtered new phrases in different languages, new associations are built computing the similarity between each element of the sets. Our similarity does not take into account the word forms but the morphosyntactic information of each token of the phrase. New associations are added to the phrase and reordering models multiplying the probabilities of the original association by the similarity score: most reliable associations get the highest scores. We test the expanded models on En-Fr and Fr-En translations using two different test sets and results show improvements of the performance in terms of Bleu [18], Meteor [15] and OOV word reduction and better translation of known phrases.

This paper is structured as follows: section II reports previous work, section III describes our expansion method, section IV sets the experimental framework, section V presents the results and, finally, section VI concludes and discusses future work.

## II. RELATED WORK

A large number of work has recently been proposed to increase the knowledge of an SMT system using external resources.

Manuscript received November 2, 2010. Manuscript accepted for publication January 14, 2011.

The authors are with the Joint Research Centre (JRC), IPSC - GlobSec, European Commission, Via Fermi 2749, 21027, Ispra (VA), Italy (e-mail: name.surname@jrc.ec.europa.eu.)

A classical approach consists of adding parallel data. In [20], the authors study the translation capability of a PBSMT system under different conditions, showing that the performance does not necessary improve when adding independent and identically distributed parallel data. They also suggest the generation of artificial training data based on existing training data, or *a posteriori* expansion of the tables. We follow these suggestions in our work. Other kind of parallel data can be used: in [19], parallel treebank data are added to a PBSMT system trained with Europarl data. Different approaches to incorporate such new data are proposed. They show that it is possible to raise the translation performance but, increasing the Europarl seed, the contribution of the treebank data decreases.

The knowledge of a PBSMT system can also be increased extracting different types of information from the training data and using all of them together. Koehn and Hoang [13] integrate additional annotations at the word level such as lemma, part-of-speech and morphological features. The proposed method outperforms the baseline in terms of automatic score and grammatical coherence.

Another approach consists in using some external data (monolingual or multilingual) to increase the existing knowledge; several methods have been proposed. Our selection may be representative but not exhaustive. Marton *et al.* [16] investigate how to augment training data by deriving monolingual paraphrases that are similar (in terms of distributional profiles) to OOV words and phrases, using distributional semantic similarity measures. Mirkin *et al.* [17] also propose an entailment-based approach to handle unknown words, using a source-language monolingual resource (WordNet) and a set of textual entailment rules. Both approaches show better results compared to the baseline. Haffari *et al.* [9] propose an active learning framework and try several sentence selection strategies, showing results accordingly. In [6], Garcia *et al.* propose to use a multilingual lexical database to compute more informed translation probabilities, showing good results when applying the MT system to a new domain.

Regarding the use of morphology in the SMT, a lot of work has been done (see Yang and Kirchhoff [21]), but few of it has analysed directly the phrase table content. When encountering unseen verbal forms, De Gispert *et al.* [3] look for similar known forms and generate new phrases on the source and target sides, using morphological and shallow syntax information. With this method, they show improvements in terms of Bleu score. Yang and Kirchhoff [21] propose a hierarchical backoff model based on morphological information: for an unseen word, the model relies on translation probabilities derived from stemmed or split versions of the word. Habash [8] uses morphological inflection rules to match OOV words with INV (in vocabulary) words and to generate new phrases in which INV words are replaced by OOV words. In his experiments, this approach allows the system to handle 60% of the OOV.

In this paper, we propose a morphologically-based method to expand the existing knowledge of an SMT system. This new knowledge is then used by the PBSMT system to handle unseen words and to produce more reliable translations for seen words. As far as we know, this is the first attempt to generate new high quality associations using morphological resources and considering *all* original associations in the phrase table, whatever their part of speech is.

## III. KNOWLEDGE EXPANSION

In this work, we focus our attention on the fact that, in an SMT system, each word form is treated as a token: two words, one morphological variation of the other, are different and independent tokens. Therefore, if one of the morphologically-related word forms is not in the training data, the word will become an OOV word or will be wrongly translated. Let's consider an example, from French to English:
**SOURCE**: *. . . les élections parlementaires anticipées en autriche ont apporté un affaiblissement sensible de la principale coalition . . .*
**TARGET**: *. . . the early parliamentary elections in austria have apporté*|||UNK *a weakening sensitive of the principal coalition* . . .
In the translated sentence, the word *apporté* is not translated (marked as unknown) and the word *principale* is translated as *principal* instead of *leading* (as it is in the reference sentence), even if in the translation phrase table learned during the training phase we have the following associations[1]:
`apporte ||| brings` ### `apporte ||| provides` ### `nous apportons ||| we provide` ### `principale ||| principal` ### `principales ||| leading`

Our approach proposes to use morphological resources to expand the knowledge of the system: new associations are generated and added to the phrase and reordering models; these new associations contain morphological variations of source and target phrases created during the training process. Regarding the previous example, the phrase table (PT) will be expanded with the associations `apporté ||| brought` and `principale|||leading`, enabling the SMT system to correctly translate the sentence.

The process of generation of new associations takes as input the phrase and reordering tables on one side, and morphological resources on the other. In our experiments we used the English and French Multext morphological resources [4]. These morphosyntactic lexicons provide, for each lexical entry, three types of information: the word form (*brought*), its lemma (*bring*), and finally its MorphoSyntactic Description (MSD, *Vviq3s*). The MSD is a condensed tag that encodes the morphosyntactic features of the word, in the form of attribute-value pairs specified via letters (part of speech, gender, number, tense, mood, etc.). One significant advantage of Multext resources is that they provide harmonized morphosyntactic description for more than 15 languages. The whole chain of knowledge expansion is made up of five steps, described in the next sections.

---

[1] Only the textual part is presented here.

**Monolingual Expansion of a Phrase.** Given an association from the PT, the first objective is to generate all the possible morphological variations for each of its monolingual parts. For each token of a monolingual phrase, we first generate a vector that contains all its morphological variations. To do so, we look for its associated lemma(s) in the morphological resources and return all the words that share this lemma. We then apply a recursive algorithm that takes, for each phrase, the morphological variation vector and produces new phrases, in which each token is associated with its MSD. This monolingual expansion phase is done for all the tokens, whatever their part of speech (POS) is. In our example, if we take the phrase `nous apportons`, we first expand “nous” then “apportons” and finally we build new phrases, as illustrated in Figure 1. Due to the absence of any constraints in this phrase expansion step, wrong phrases are generated (marked with stars in Figure 1). A filtering step is therefore needed.

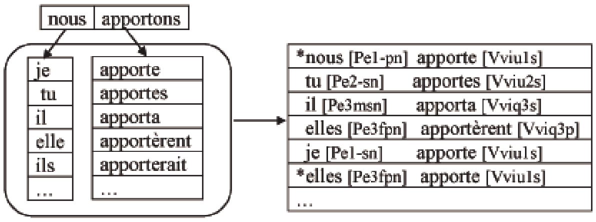
nous
apportons
je
tu
il
elle
ils
...
apporte
apportes
apporta
apportèrent
apporterait
...
*nous [Pe1-pn] apporte [Vviu1s]
tu [Pe2-sn] apportes [Vviu2s]
il [Pe3msn] apporta [Vviq3s]
elles [Pe3fpn] apportèrent [Vviq3p]
je [Pe1-sn] apporte [Vviu1s]
*elles [Pe3fpn] apporte [Vviu1s]
...

Fig. 1. Example of a monolingual phrase expansion.

**Phrase Filtering.** We defined two kinds of filtering. The first one, based on probability checking, is designed to carry out a coarse selection. The second one, based on grammatical rules, performs a fine-grained selection.

Probability checking filtering takes advantage of the language models created using more than 3 million sentence pairs. For each language, the language model is queried with the generated phrases and then probabilities of correctness are computed for each phrase. The list of phrases is sorted according to the probabilities and only phrases above a defined threshold are kept. This threshold was computed using human-annotated data: given a randomly selected set of phrases for each phrase length (from 1 to 7 tokens) and for each language, phrases were expanded and manually annotated according to their grammatical and semantic correctness. Thousands of new phrases were annotated for English and French. These phrases were then sorted according to their probabilities (computed against the LM) and, for each possible threshold value, the $F_{0.5}$ score was calculated in reference to annotated data. We used the $F_{0.5}$ score because it weights precision twice as much as recall, and we prefer to generate good quality data, even if there are less new associations. For each language and phrase length, we computed the maximum $F_{0.5}$ score values and took their relative threshold values. Figure 2 illustrates the threshold computation for English phrase lengths.

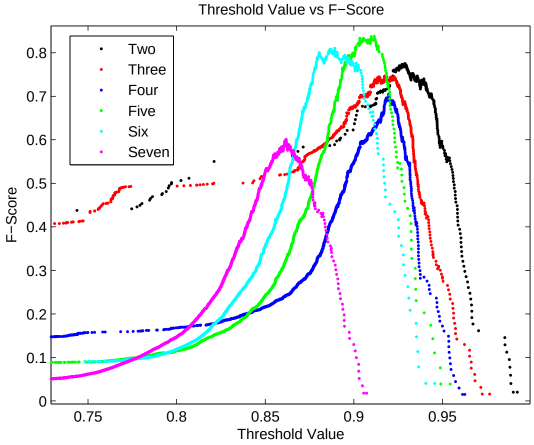

Fig. 2. Computation of the English thresholds by phrase length.

This language model filtering approach is able to remove an important number of wrong phrases; however, it is data-dependent and does not filter out all the unwanted phrases. To supplement this first coarse filtering, we use a set of hand-written grammatical rules based on morphosyntactic tags. The rules allow to verify gender and number agreement between tokens, to check tense agreements within verbal phrases, and/or to remove obvious wrong sequences of tags (like three or four times the same POS tag in a row). This second filtering gives, again, the opportunity to remove wrong phrases from the list. At this stage we have, for each language, a list of correct new phrases (in Figure 1, only phrases without stars remain in the list). These new phrase lists are then used to produce new associations.

**Generation of New Associations.** The objective of this step is, given two sets of new phrases, to create new associations. To match phrases, we use a string matching similarity score based on morphosyntactic information. If we consider the PT association `nous apportons`|||`we bring`|||`(0)(1)`|||`(0)(1)` (numbers state the word alignment), first steps should have produced two lists of correct new phrases, where each token is associated with its MSD, ( `il[Pe3msn] apporta[Vviq3s]` and `we[Pe1-pn] brought[Vviq2p]` ).

Given two phrases ($p_1$ and $p_2$) in different languages, the morphosyntactic descriptions of their tokens ($t_1^{msd}$ and $t_2^{msd}$) and the number of elements in the word alignment ($a$) of the original association, we compute the similarity as:

$$s(p_1, p_2) = \frac{\sum\limits_{i,j \in a} st(t_i^{msd}, t_j^{msd})}{|a|} \tag{1}$$

TABLE I
EXAMPLES OF MSD SIMILARITY SCORE COMPUTATION

| | **Vviq3s** and **Vviq2p** | | | | | | **Pe3msn** and **Pe1-pn** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MSD1 | v | i | q | 3 | s | | e | 3 | m | s | n | |
| MSD2 | v | i | q | 2 | p | | e | 1 | - | p | n | |
| Score | 1 | 1 | 1 | 0 | 0 | 3/5 | 1 | 0 | 0,5 | 0 | 1 | 2.5/5 |

TABLE II
MANUAL EVALUATION OF NEW ASSOCIATIONS GENERATED EXPANDING 1,000 RANDOM PT ENTRIES

| Alignment Type | Precision | Number of New Associations |
|---|---|---|
| **A** | 0.6725 | 1933 |
| **no M** | 0.7544 | 1820 |
| **no M + E** | 0.8261 | 1530 |
| **no M + E + OE** | **0.8861** | 1115 |

TABLE III
NUMBER OF ENTRIES IN THE PHRASE TABLE

| | Fr-En (News) | En-Fr (News) | En-Fr (Europ.) |
|---|---|---|---|
| **Original** | 3,946,143 | 3,924,804 | 60,873,395 |
| **Reduced** | 229,390 | 217,685 | 4,480,135 |
| **Expanded** | 345,896 | 334,188 | 5,671,418 |

where:

$$st(t_i^{msd}, t_j^{msd}) = \frac{\sum_{i \in len(t_i^{msd})} m(t_i^{msd}(i), t_j^{msd}(j))}{len(t_i^{msd})} \qquad (2)$$

The similarity between two phrases corresponds to the sum of the similarities between two tokens, normalized by the numbers of aligned tokens in the original associations (1); then, the similarity between two tokens corresponds to the similarity between two morphosyntactic descriptions given a matrix *m*, normalized by the length of the MSD (2). Considering the two new phrases generated from the PT association, `il[Pe3msn] apporta[Vviq3s]` and `we[Pe1-pn] brought[Vviq2p]`, the similarity between these phrases is equal to the similarity between the MSD "Pe3msn" (from *il*) and the MSD "Pe1-pn" (from *we*) plus the similarity between the MSD "Vviq3s" (from *apporta*) and the MSD "Vviq2p" (from *brought*), all divided by 2, which corresponds to the number of elements in the original association alignment (`(0)(1)`|||`(0)(1)`). In case of multi-alignment, the similarity of the single token is computed against all its aligned tokens.

The similarity between MSDs corresponds to a positional score based on a substitution matrix: each entry in the matrix describes the rate at which one character (in our case a letter encoding morphosyntactic information) in a MSD can be changed to another. Matrices were manually built by a linguist for the following parts of speech: Noun, Verb, Adjective, Pronoun, Determiner, Adverb, Preposition, Conjunction and Numeral. Within matrices, we decided to use the following values: 0 for morphological information that should not be matched (singular with plural for example), 0.5 for information that can be matched but not necessarily (feminine with neutral) and 1 when information should be matched (present tense with present tense). Regarding our example, the similarities between `apporta[Vviq3s] brought[Vviq2p]` and `il[Pe3msn] we[Pe1-pn]` are illustrated in Table I. Single character scores are obtained querying the Verb and Pronoun matrices (V and P).

For all potential phrase associations from the filtered lists of expanded phrases, we computed the similarity as described above. We then ranked the associations by similarity and computed a threshold corresponding to: $max - (max * 10\%)$, *max* being the maximum similarity value of the new association set. We finally keep the associations which have similarity values bigger than this threshold. In our example, the similarity between the two phrases is $\frac{\frac{2.5}{5}+\frac{3}{5}}{2} = 0.55$. If we have the same MSDs in both phrases, the maximum reachable would be 1 and the relative threshold is 0.9. In this case, the new association would be discarded. At the end, we have at our disposal "artificial" new associations that can be added to the phrase and reordering tables. Before doing so, we completed an evaluation of the new associations.

**New Association evaluation.** To evaluate the new associations, we randomly selected 1,000 associations from the Fr-En phrase table, expanded them using our algorithm and manually annotated.[2] The manual annotation was done in rather a strict way: an association was considered as correct if there was no mistake, neither in the phrases, nor in the association. Regarding the original association we did not judge its quality but we took into account different types of alignment. We distinguish between the following cases: *multi-alignment* (M), when a token on one side is aligned with several on the other ((0)(0)(0)… | (0,1,2)…), *one empty alignment* (OE), when one token on one side does not have a correspondence on the other ((0)() | (0)), and *several empty alignments* (E), when more than one token on one side does not have corresponding tokens on the other (()(1)()(0)()()() | (3)(1)). We computed the Precision according to these different types.

Results are presented in Table II. Precision is affected by two phenomena: the type of alignment taken into account and the phrase length (results by phrase length are omitted due to lack of space). Essentially, the measure increases removing multi and empty alignments (we add less new associations but of better quality), and considering shorter phrases. Showing up cases where new associations are of better or lower quality, this evaluation helped us to decide which type of original association to expand. The next section considers how to add new associations to the model.

**Integration of New Associations.** Starting from the PT, we artificially generate new associations that are finally added to the phrase and reordering models which constitute, at the end, an extended model. While adding new data to the original tables, we pay attention to do so respecting the way the data were produced. New associations are made of three parts: a textual part (the newly generated phrases), a "word order" part (we keep the same as the original association), and a probability part, that indicates how reliable an association is. Probabilities taken in account are bidirectional translation probabilities and lexical weighing for the phrase table and bidirectional monotone, swapped and discontinuous reordering probabilities in the reordering model. In our extended model, the probability of a new association is computed multiplying the probabilities of the original association by the similarity score of the new association. This allows two things: at the phrase table level, original associations get the highest probabilities; then, within a set of new associations generated from a particular PT association, each new association has its own probability, reflecting how reliable its generation process was. In this way, phrase and reordering tables can be extended without perturbing the original knowledge of the system. Finally, if a new association is a duplicate of an original one, it is not added to the new model.

---

[2] As the expansion process is symmetric, the evaluation from the fr-en phrase table is also valid for the en-fr one.

## IV. EXPERIMENTAL SETTING

To assess our approach, we conducted a series of experiments on French-to-English and English-to-French translations. They were run using Moses [14], a complete phrase-based machine translation toolkit for academic purposes, and IRSTLM [5] for language modelling during the phrase filtering and the pure translation. Results have been evaluated in terms of Bleu and Meteor scores over lowercased output and number of OOV words. Meteor considers the surface form of each word and does not make use of WordNet synonyms.

**SMT data.** We trained the PBSMT models using two different training corpora: Commentary English-French and French-English News corpus [1] containing 64,233 sentence pairs in both directions and Europarl Release v3 English-French [12] containing 1,428,799 sentence pairs. We used two test sets in both language pair directions coming from different domains: 3,000 sentences from Commentary News and 2,000 from the proceedings of Europarl, both selected by the organizers of the Statistical Machine Translation Workshop [1].

**Pre-Processing of the PT.** The translation table contains all phrase pairs found in the training corpus, which includes a lot of noise. Our approach expands all the associations found in the translation model without taking into account their correctness. To avoid the expansion of unreliable associations, we pre-processed the phrase table using the method proposed by Johnson et al. [10]. This approach prunes the PT using a technique based on the significance testing of phrase pair co-occurrence in the parallel corpus. In our experiments we used a threshold equal to $\alpha + \epsilon$ and only the top 30 phrase translations for each source phrase based on $p(e|f)$ were kept. If a phrase pair appears exactly once in the corpus and each of the component phrases occurs exactly once on its side of the parallel corpus, this special case is called 1-1-1 association. Our parameter choices removed all the 1-1-1 associations.

New PT dimensions resulting from this pruning process are shown in Table III. Testing the original and reduced models on the test data confirmed the results found in [10]: substantial reduction of the PT dimension does not alter the translation performance. In the rest of the paper, baseline results refer to the performance of the reduced model.

## V. RESULTS

Three PBSMT systems were built using the training data presented above: French-English translation trained on the Commentary News data ($F2E_N$), English-French translation trained on the Commentary News data ($E2F_N$) and English-French translation trained on the Europarl data ($E2F_E$). For each of them, phrase tables were pre-processed and then expanded using our algorithm. According to what we learnt from the evaluation of the new associations (section III), we expanded original associations that do not contain multi or empty (one or more than one) alignments. Even if this choice reduced the number of new associations, it guarantees high precision in what is added (Table II).

The expansion of phrase and reordering models requires a counterpart information in the language model. Thus, a language model was created using the target side of the training data plus an external corpus crawled on the Web containing 3,463,954 French and 3,183,871 English sentences. Performance of the baseline and extended models is shown on the left side of Table IV. In all experiments, the number of OOV words decreases; this is more evident in Fr-En translation, as the source language is more morphologically inflected. In terms of automatic scores, the $F2E_N$ and $E2F_N$ expanded models resulted in improvements with respect to the baseline.

Knowledge expansion should allow the model not only to translate unknown words (in our initial example, *apporté* is translated into *brought*) but also to better translate already known ones (*principal* is replaced by *leading*). In order to evaluate this phenomenon, we conducted a manual evaluation on a set of 110 randomly selected target sentences ($F2E_N$) where there is a difference (increase or decrease) in Meteor score between the baseline and expanded system translations. Comparing them, we distinguished several causes of score variation: unknown word covered (Unknown), known word substituted (Known), unknown word covered and known word substituted (Both) and other reasons like word reordering (Other). The results of this manual evaluation (Table V) confirm that the expanded model performs better than the baseline and show that improvements not only come from unknown word coverage but also from better translations of known words.

From a manual analysis of the $F2E_N$ translated sentences, we additionally noticed that in several cases the automatic scores are not able to capture improvements given by the expanded models, see [2] for more details on this problem.

TABLE IV
OBTAINED RESULTS

| | 3-gram Language Model | | | | 2-gram Language Model (test set) | | | |
|---|---|---|---|---|---|---|---|---|
| | Commentary News | | Europarl | | Commentary News | | Europarl | |
| **Fr-En Commentary News ($F2E_N$)** | | | | | | | | |
| | Baseline | Expanded | Baseline | Expanded | Baseline | Expanded | Baseline | Expanded |
| **Bleu %** | 21.68 | 21.89 | 21.99 | **22.37** * | 26.41 | 27.01 * | 26.46 | **27.17** * |
| **Meteor** | 0.4698 | **0.4733** * | 0.4706 | 0.4720 | 0.4975 | 0.5035 * | 0.4972 | **0.5042** * |
| **OOV** | 7,763 | **7,004** | 3,107 | 2,741 | 7,763 | **7,004** | 3,107 | 2,741 |
| **En-Fr Commentary News ($E2F_N$)** | | | | | | | | |
| **Bleu %** | 21.35 | 21.61 * | 23.62 | **23.79** * | 24.66 | **25.22** * | 25.89 | 26.36 * |
| **Meteor** | 0.1524 | 0.1542 * | 0.1630 | **0.1650** * | 0.1739 | 0.1780 * | 0.1805 | 0.1842 * |
| **OOV** | 6,447 | **5,977** | 2,400 | 2,153 | 6,447 | **5,977** | 2,400 | 2,153 |
| **En-Fr Europarl ($E2F_E$)** | | | | | | | | |
| **Bleu %** | 22.62 | **22.63** | 27.43 | 27.38 | 28.51 | **28.73** * | 34.75 | 34.77 |
| **Meteor** | 0.1608 | 0.1607 | 0.1927 | 0.1923 | 0.2025 | 0.2040 * | 0.2465 | 0.2467 |
| **OOV** | 3,357 | **3,186** | 260 | 253 | 3,357 | **3,186** | 260 | 253 |

TABLE V
HUMAN EVALUATION OF A SAMPLE OF 110 RANDOM SELECTED SENTENCES FROM $E2F_N$

| | Total | Unknown | Known | Both | Other |
|---|---|---|---|---|---|
| **Increment in Meteor** | 84 | 18 | 45 | 2 | 19 |
| | | 21.4% | 53.5% | 2.3% | 22.6% |
| **Decrement in Meteor** | 26 | 0 | 16 | 0 | 10 |
| | | 0 | 61.5% | 0 | 38.5% |

**Baseline:** *we have settled our divergentes|||UNK views …*
**Expanded:** *we have settled our divergent views …*
**Reference:** *we've resolved our differing opinions …*

In this example, the expanded sentence has no OOV words and is more comprehensible for a non-French speaker, but there is not improvement regarding the automatic scores. This kind of example, combined with the need of a counterpart in the language model, raised the following question: Was the correct translation of the word *divergentes* – according to the reference sentence – present in the model?

**Controlled environment experiments.** To answer these questions, we ran a set of controlled environment experiments. Our idea was to evaluate only the knowledge of the phrase and reordering models cutting out the language model contribution. Instead of using the big language model, which obviously was not exhaustive and could negatively influence the performance, we used a 2-gram language model built on the target side of the test set. Regardless of the small number of sentences used and of the fact that probabilities may not be accurately estimated, it drove the decoder to select those phrases that were present in the reference sentences. Differences in performance between the baseline and expanded models reflect only the difference in terms of knowledge in the phrase and reordering tables. Results are shown on the right side of Table IV.

Results in the Table IV are obtained using a 3-gram language model trained on the target side of the training data plus 3,463,954 French sentences or 3,183,871 English sentences. * = significance test over baseline with $p < 0.0001$, using pair-wise bootstrap test with 95% confidence interval [11]

In these controlled environment experiments, the gap between the baseline and the expanded models increased with a maximum 0.73 Blue score points. The augmented system has a significant gain over its baseline also in the $E2F_E$ translations using the out-of-domain test set. These results show how the new model took advantage of the information added by the new associations, increasing the quality of the output translations. This means that the new model has the correct information to produce a target sentence similar to the reference sentence, but the selection of the correct translation option is strictly related to the language model information. Target sentences that are not similar to the reference sentences are not necessarily wrong.

## VI. DISCUSSION AND FUTURE WORK

This work shows that the knowledge of a Statistical Machine Translation system can be artificially expanded without relying on parallel data. Morphological resources are used to generate new high quality associations that are added to phrase and reordering models. Each new association contains source/target/both phrases that are morphological variations of the original ones. Although this may be considered a limitation, because "never seen" associations cannot be added, results confirm the benefits in terms of translation quality.

Our algorithm increases the dimension of the PTs (see Table III): for models trained with Commentary News roughly about 50%, while for the Europarl model about 25%. This assumes particular relevance if we thought that in the reduced tables 1-1-1 associations are pruned, see Section IV. It means that each new association that the proposed method adds would require at least more than one parallel sentence pairs to be added during the training phase using parallel data.

Empirical results support the assumption that the new associations help the SMT system to better translate sentences coming from different domains. Our expanded models performed better than the baseline in particular when the original model is trained on a small training set. It reduces the impact of the OOV words in the translation, but not only:

manual evaluation shows that also known words are replaced by better translations in the target sentences.

Manual analysis of the results highlighted the weakness of the automatic score to catch improvements in translation. This suggested to us a series of experiments with a small but optimal language model. In this controlled environment, results show even more benefits of our approach with any different training set sizes and test data. This confirms that extra knowledge can be used by the decoder only with a language model that contains suitable information.

Our intention is to make our technique more portable to other language pairs replacing the grammatical rules with a language model built on part of speech information. The idea of expanding the knowledge of an SMT system is generic and different types of information can be passed artificially to it. In this paper we investigated how to add morphologically related new associations; in a next step, we will consider how to add new semantically related associations, e.g. semantic knowledge. We believe that the benefits of our approach will be more evident using more inflected languages like Czech. Experiments are planed in this direction.

## ACKNOWLEDGEMENT

The authors would like to thank Nello Cristianini for preliminary discussion about this work, Ralf Steinberger and Jenya Belyaeva for comments and suggestions that helped to improve the present paper.

## REFERENCES

[1] C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder, "Findings of the 2009 Workshop on Statistical Machine Translation," in *Proceedings of WSMT*, 2009, pp. 1-28.

[2] C. Callison-Burch and M. Osborne, "Re-evaluating the role of BLEU in machine translation research," in *Proceedings of EACL*, 2006, pp. 249-256.

[3] A. De Gispert, J.B. Mariño, and J.M. Crego, "Improving statistical machine translation by classifying and generalizing inflected verb forms," in *Proceedings of 9th European Conference on Speech Communication and Technology*, 2005, pp. 3193-3196.

[4] T. Erjavec, "MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora," in *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation,* 2004.

[5] M. Federico, N. Bertoldi, and M. Cettolo, "Irstlm: an open source toolkit for handling large scale language models," in *Proceedings of Interspeech,* 2008, pp. 1618-1621.

[6] M. Garcia, J. Giménez, and L. Màrquez, "Enriching Statistical Translation Models Using a Domain-Independent Multilingual Lexical Knowledge Base," *Lecture notes in computer science (Computational Linguistics and Intelligent Text Processing),* vol. 5449, pp. 306-317, 2009.

[7] S. Goldwater and D. McClosky, "Improving statistical MT through morphological analysis," in *Proceedings of EMNLP*, 2006, pp. 676-683.

[8] N. Habash, "Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation," in *Proceedings of ACL*, 2006, pp. 57-60.

[9] G. Haffari, M. Roy, and A. Sarkar, "Active learning for statistical phrase-based machine translation," in *Proceedings of NAACL*, 2009, pp. 415-423.

[10] H. Johnson, J. Martin, G. Foster, and R. Kuhn, "Improving translation quality by discarding most of the phrasetable," *Proceedings of EMNLP-CoNLL*, 2007, pp. 967-975.

[11] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of EMNLP*, 2005, pp. 388-395.

[12] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of MT summit*, 2005.

[13] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of EMNLP-CoNLL*, 2007, pp. 868-876.

[14] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico and others, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of ACL, demonstration session*, 2007, pp. 1618-1621.

[15] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, pp. 228-231.

[16] Y. Marton, C. Callison-Burch, and P. Resnik, "Improved statistical machine translation using monolingually-derived paraphrases," in *Proceedings of EMNLP*, 2009, pp. 381-390.

[17] S. Mirkin, L. Specia, N. Cancedda, I. Dagan, M. Dymetman, and I. Szpektor, "ource-language entailment modeling for translating unknown terms," in *Proceedings of ACL*, 2009, pp. 791-799.

[18] K. Papineni, S. Roukos, T. Ward and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, 2002, pp. 311-318.

[19] J. Tinsley, M. Hearne and A. Way, "Exploiting parallel treebanks to improve phrase-based statistical machine translation," in *Proceedings of CICLing*, 2009, pp. 318-331.

[20] M. Turchi, T. DeBie, and N. Cristianini, "Learning performance of a machine translation system: a statistical and computational analysis," *Proceedings of the Third Workshop on Statistical Machine Translation*, 2008, pp. 35-43.

[21] M. Yang and K. Kirchhoff, "Phrase-based backoff models for machine translation of highly inflected languages," in *Proceedings of EACL*, 2006, pp. 41-48.