**Improving Reordering with Linguistically Informed Bilingual** *n***-grams**

**Josep Maria Crego** LIMSI-CNRS [jmcrego@limsi.fr](mailto:jmcrego@limsi.fr)

**Franc¸ois Yvon**

LIMSI-CNRS & Univ. Paris Sud

[yvon@limsi.fr](mailto:yvon@limsi.fr)

**Abstract**

We present a new reordering model es- timated as a standard *n*-gram language model with units built from morpho- syntactic information of the source and target languages. It can be seen as a model that translates the morpho-syntactic structure of the input sentence, in contrast to standard translation models which take care of the surface word forms. We take advantage from the fact that such units are less sparse than standard translation units to increase the size of bilingual con- text that is considered during the trans- lation process, thus effectively account- ing for mid-range reorderings. Empirical results on French-English and German- English translation tasks show that our model achieves higher translation accu- racy levels than those obtained with the widely used lexicalized reordering model.

**1 Introduction**

Word ordering is one of the major issues in statis- tical machine translation (SMT), due to the many word order peculiarities of each language. It is widely accepted that there is a need for struc- tural information to account for such differences. Structural information, such as Part-of-speech (POS) tags, chunks or constituency/dependency parse trees, offers a greater potential to learn generalizations about relationships between lan- guages than models based on word surface forms, because such “surfacist” models fail to infer gen- eralizations from the training data.

The word ordering problem is typically decom- posed in a number of related problems which can be further explained by a variety of linguistic phe- nomena. Accordingly, we can sort out the re- ordering problems into three categories based on

the kind of linguistic units involved and/or the typical distortion distance they imply. Roughly speaking, we face *short-range* reorderings when single words are reordered within a relatively small window distance. It consist of the easi- est case as typically, the use of phrases (in the sense of translation units of the phrase-based ap- proach to SMT) is believed to adequately perform such reorderings. *Mid-range* reorderings involve reorderings between two or more phrases (trans- lation units) which are closely positioned, typi- cally within a window of about 6 words. Many alternatives have been proposed to tackle mid- range reorderings through the introduction of lin- guistic information in MT systems. To the best of our knowledge, the authors of (Xia and Mc- Cord, 2004) were the first to address this prob- lem in the statistical MT paradigm. They auto- matically build a set of linguistically grounded rewrite rules, aimed at reordering the source sen- tence so as to match the word order of the target side. Similarly, (Collins, et al 2005) and (Popovic and Ney, 2006) reorder the source sentence us- ing a small set of hand-crafted rules for German- English translation. (Crego and Marin˜ o, 2007) show that the ordering problem can be more accu- rately solved by building a source-sentence word lattice containing the most promising reordering hypotheses, allowing the decoder to decide for the best word order hypothesis. Word lattices are built by means of rewrite rules operating on POS tags; such rules are automatically extracted from the training bi-text. (Zhang, et al 2007) introduce shallow parse (chunk) information to reorder the source sentence, aiming at extending the scope of their rewrite rules, encoding reordering hypothe- ses in the form of a confusion network that is then passed to the decoder. These studies tackle mid-range reorderings by predicting more or less accurate reordering hypotheses. However, none

197

*Coling 2010: Poster Volume*, pages 197–205, Beijing, August 2010

of them introduce a reordering model to be used in decoding time. Nowadays, most of SMT sys- tems implement the well known *lexicalized re- ordering* model (Tillman, 2004). Basically, for each translation unit it estimates the probability of being translated *monotone*, *swapped* or placed *discontiguous* with respect to its previous trans- lation unit. Integrated within the *Moses* (Koehn, et al 2007) decoder, the model achieves state-of- the-art results for many translation tasks. One of the main reasons that explains the success of the model is that it considers information of the source- and target-side surface forms, while the above mentionned approaches attempt to hypoth- esize reorderings relying only on the information contained on the source-side words.

Finally, *long-range* reorderings imply reorder- ings in the structure of the sentence. Such re- orderings are necessary to model the translation for pairs like Arabic-English, as English typically follows the SVO order, while Arabic sentences have different structures. Even if several attempts exist which follow the above idea of making the ordering of the source sentence similar to the tar- get sentence before decoding (Niehues and Kolss,

2009), long-range reorderings are typically better addressed by syntax-based and hierarchical (Chi- ang, 2007) models. In (Zollmann et al., 2008),

ables to account for arbitrary large sequences of units. Hence, the proposed model takes care of the translation adequacy of the structural informa- tion present in translation hypotheses, here intro- duced in the form of POS tags. We also show how the new model compares to a widely used *lexical- ized reordering* model, which we have also im- plemented in our particular bilingual *n*-gram ap- proach to SMT, as well as to the widely known *Moses* SMT decoder, a state-of-the-art decoder performing lexicalized reordering.

The remaining of this paper is as follows. In Section 2 we briefly describe the bilingual *n*-gram SMT system. Section 3 details the bilingual *n*- gram reordering model, the main contribution of this paper, and introduces additional well known reordering models. In Section 4, we analyze the reordering needs of the language pairs considered in this work and we carry out evaluation experi- ments. Finally, we conclude and outline further work in Section 5.

**2 Bilingual** *n***-gram SMT**

Our SMT system defines a translation hypothesis *t* given a source sentence *s*, as the sentence which maximizes a linear combination of feature func- tions:

an interesting comparison between phrase-based,

hierarchical and syntax-augmented models is car-

( *M*

1 = arg max

*t*ˆ*I*

*t*

)

*λmhm*(*sJ I*

(1)

ried out, concluding that hierarchical and syntax-

*I*

1 *m*=1

1 *, t*1 )

based models slightly outperform phrase-based models under large data conditions and for suf- ficiently non-monotonic language pairs.

Encouraged by the work reported in (Hoang and Koehn, 2009), we tackle the mid-range re- ordering problem in SMT by introducing a *n*- gram language model of bilingual units built from POS information. The rationale behind such a model is double: on the one hand we aim at in- troducing morpho-syntactic information into the reordering model, as we believe it plays an im-

where *λm* is the weight associated with the fea- ture *hm*(*s, t*). The main feature is the log-score of the translation model based on bilingual *n*-grams. This model constitutes a language model of a par- ticular *bi-language* composed of bilingual units which are typically referred to as *tuples* (Marin˜ o et al., 2006). In this way, the translation model prob- abilities at the sentence level are approximated by using *n*-grams of tuples:

*K*

*p*(*sJ I* n

portant role for predicting systematic word or- dering differences between language pairs; at the

1

1 *, t*1) =

*k*=1

*p*((*s, t*)*k |*(*s, t*)*k−*1 *. . .* (*s, t*)*k−n*+1)

same time that it drastically reduces the sparse- ness problem of standard translation units built from surface forms. On the other hand, *n*-gram language modeling is a robust approach, that en-

*s*

1

where *s* refers to source *t* to target and (*s, t*)*k* to the *kth* tuple of the given bilingual sentence pairs, *J* and *tI* . It is important to notice that, since both languages are linked up in tuples, the context

information provided by this translation model is bilingual. As for any standard *n*-gram language model, our translation model is estimated over a training corpus composed of sentences of the lan- guage being modeled, in this case, sentences of the *bi-language* previously introduced. Transla- tion units consist of the core elements of any SMT system. In our case, tuples are extracted from a word aligned corpus in such a way that a unique segmentation of the bilingual corpus is achieved, allowing to estimate the *n*-gram model. Figure 1 presents a simple example illustrating the unique tuple segmentation for a given word-aligned pair of sentences (top).



alignments.

Following on the previous example, the rule

*perfect translations ❀ translations perfect* pro-

duces the swap of the English words that is ob-

served for the French and English pair. Typically, POS information is used to increase the general- ization power of such rules. Hence, rewrite rules are built using POS instead of surface word forms. See (Crego and Marin˜ o, 2007) for details on tuples extraction and reordering rules.

**3 Reordering Models**

In this section, we detail three different reordering models implemented in our SMT system. As pre- viously outlined, the purpose of reordering mod- els is to accurately learn generalizations for the word order modifications introduced on the source side during the tuple extraction process.

**3.1 Source** *n***-gram Language Model**

We employ a *n*-gram language model estimated over the source words of the training corpus af- ter being reordered in the tuple extraction process. Therefore, the model scores a given source-side reordering hypothesis according to the reorder-



ings performed in the training sentences.



POS tags are used instead of surface forms

Figure 1: *Tuple extraction from an aligned sen-*

*tence pair.*

The resulting sequence of tuples *(1)* is further refined to avoid *NULL* words in source side of the tuples *(2)*. Once the whole bilingual training data

in order to improve generalization and to reduce sparseness. The model is estimated as any stan- dard *n*-gram language model, described by the following equation:

*J*

is segmented into tuples, *n*-gram language model

*p*(*sJ I*

n

*j j* 1

*|*

*j n*+1

probabilities can be estimated. Notice from the

*j*

1 *, t*1) =

*j*=1

*p*(*st st*

*−*

*, . . . , st*

*−*

) (2)

example that the English source words *perfect* and

where *st*

relates to the POS tag used for the *jth*

*translations* have been reordered in the final tu- ple segmentation, while the French target words are kept in their original order. During decoding, sentences to be translated are encoded in the form of word lattices containing the most promising re- ordering hypotheses, so as to reproduce the word order modifications introduced during the tuple extraction process. Hence, at decoding time, only those reordering hypotheses encoded in the word lattice are examined. Reordering hypotheses are introduced following a set of reordering rules au- tomatically learned from the bi-text corpus word

source word.

The main drawback of this model is the lack of knowledge of the hypotheses on the target- side. The probability assigned to a sequence of source words is only conditioned to the sequence of source words.

**3.2 Lexicalized Reordering Model**

A broadly used reordering model for phrase-based systems is lexicalized reordering (Tillman, 2004). It introduces a probability distribution for each phrase pair that indicates the likelihood of being

translated *monotone*, *swapped* or placed *discon- tiguous* to its previous phrase. The ordering of the next phrase with respect to the current phrase is typically also modeled. In our implementa- tion, we modified the three orientation types and

Figure 2 shows the sequence of translation units built from POS tags, used in our previous exam- ple.

consider: a *consecutive* type, where the original

monotone and swap orientations are lumped to- gether, a *forward* type, specifying discontiguous forward orientation, and a *backward* type, spec- ifying discontiguous backward orientation. Em- pirical results showed that in our case, the new orientations slightly outperform the original ones. This may be explained by the fact that the model is applied over tuples instead of phrases.

Counts of these three types are updated for each unit collected during the training process. Given these counts, we can learn probability dis-

tributions of the form *pr* (*orientation|*(*st*)) where

*orientation ∈ {c, f, b}* (consecutive, forward

and backward) and (*st*) is a translation unit.

Counts are typically smoothed for the estimation of the probability distribution. A major weakness of the lexicalized reordering model is due to the fact that it does not considers phrase neighboring, *i.e.* a single probability is learned for each phrase pair without considering its context. An additional concern is the problem of sparse data: translation units may occur only a few times in the training data, making it hard to estimate reliable probabil- ity distributions.

**3.3 Linguistically Informed Bilingual**

*n***-gram Language Model**

The bilingual *n*-gram LM is estimated as a stan- dard *n*-gram LM over translation units built from POS tags represented as:

*K*

Figure 2: *Sequence of POS-tagged units used to estimate the bilingual n-gram LM.*

POS-tagged units used in our model are ex- pected to be much less sparse than those built from surface forms, allowing to estimate higher order language models. Therefore, larger bilingual con- text are introduced in the translation process. This model can also be seen as a translation model of the sentence structure. It models the adequacy of translating sequences of source POS tags into tar- get POS tags.

Note that the model is not limited to using POS information. Rather, many other informa- tion sources could be used (supertags, additional morphology features, *etc.*), allowing to model dif- ferent translation properties. However, we must take into account that the degree of sparsity of the model units, which is directly related to the in- formation they contain, affects the level of bilin- gual context finally introduced in the translation process. Since more informed units may yield more accurate predictions, more informed units may also force the model to fall to lower *n*-grams. Hence, the degree of accuracy and generalization power of the model units must be carefully bal- anced to allow good reordering predictions for contexts as large as possible.

As any standard language model, smoothing is needed. Empirical results showed that Kneser- Ney smoothing (Kneser and Ney, 1995) achieved

*p*(*sJ , tI* ) = n *p*((*st*)*t |*(*st*)*t*

*. . .* (*st*)*t*

) the best performance among other options (mea-

1 1

*k*=1

*k k−*1

*k−n*+1

sured in terms of translation accuracy).

where (*st*)*t*

*k*

relates to the *kth* translation unit,

**3.4 Decoding Issues**

(*st*)*k* , built from POS tags instead of words.

This model aims at alleviating the drawbacks of

the previous two reordering models. On the one hand it takes into account bilingual information to model reordering. On the other hand it con- siders the phrase neighboring when estimating the reordering probability of a given translation unit.

A straightforward implementation of the three models is carried out by extending the log-linear combination of equation (1) with the new features. Note that no additional decoding complexity is introduced in the baseline decoding implementa- tion. Considering the bilingual *n*-gram language model, the decoder must know the POS tags for

each tuple. However, each tuple may be tagged differently, as words with same surface form may have different POS tags.

We have implemented two solutions for this sit- uation. Firstly, we assume that each tuple has a single POS-tagged version. Accordingly, we se- lect a single POS-tagged version out of the mul- tiple choices (the most frequent). Secondly, all POS-tagged versions of each tuple are allowed. The second choice implies using more accurate POS-tagged tuples to model reordering, however, it overpopulates the search space with spurious hypotheses, as multiple identical units (with dif- ferent POS tags) are considered.

Our first empirical findings showed no differ- ences in translation accuracy for both configura- tions. Hence, in the remaining of this paper we only consider the first solution (a single POS- tagged version of each tuple). The training cor- pus composed of tagged units out of which our new model is estimated is accordingly modified to contain only those tagged units considered in de- coding. Note that most of the ambiguity present in word tagging is resolved by the fact that transla- tion units may contain multiple source and target side words.

**4 Evaluation Framework**

In this section, we perform evaluation experi- ments of our novel reordering model. First, we give details of the corpora and baseline system employed in our experiments and analyze the re- ordering needs of the translation tasks, French- English and German-English (in both directions). Finally, we evaluate the performance of our model and contrast results with other reordering models and translation systems.

**4.1 Corpora**

We have used the fifth version of the *EPPS* and the *News Commentary* corpora made available in the context of the *Fifth ACL Workshop on Statistical Machine Translation*. Table 1 presents the basic statistics for the training and test data sets. Our test sets correspond to *news-test2008* and *new- stest2009* file sets, hereinafter referred to as *Tune* and *Test* respectively.

French, German and English Part-of-speech tags are computed by means of the *TreeTagger* 1 toolkit. Additional German tags are obtained us- ing the *RFTagger* 2 toolkit, which annotates text with fine-grained part-of-speech tags (Schmid and Laws, 2008) with a vocabulary of more than 700 tags containing rich morpho-syntactic information (gender, number, case, tense, *etc.*).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Lang.* | *Sent.* | *Words* | *Voc.* | *OOV* | *Refs* |
| *Train* | | | | | |
| French  English | 1*.*75 M  1*.*75 M | 52*.*4 M  47*.*4 M | 137 k  138 k | *−*  *−* | *−*  *−* |
| *Tune* | | | | | |
| French  English | 2*,* 051  2*,* 051 | 55*.*3 k  49*.*2 k | 8*,* 957  8*,* 359 | 1*,* 282  1*,* 344 | 1  1 |
| *Test* | | | | | |
| French  English | 2*,* 525  2*,* 525 | 72*.*8 k  65*.*1 k | 10*,* 832  9*,* 568 | 1*,* 749  1*,* 724 | 1  1 |
| *Train* | | | | | |
| German  English | 1*,* 61 M  1*,* 61 M | 42*.*2 M  44*.*2 M | 381 k  137 k | *−*  *−* | *−*  *−* |
| *Tune* | | | | | |
| German  English | 2*,* 051  2*,* 051 | 47*,* 8 k  49*,* 2 k | 10*,* 994  8*,* 359 | 2*,* 153  1*,* 491 | 1  1 |
| *Test* | | | | | |
| German  English | 2*,* 525  2*,* 525 | 62*,* 8 k  65*,* 1 k | 12*,* 856  9*,* 568 | 2*,* 704  1*,* 810 | 1  1 |

Table 1: *Statistics for the training, tune and test data sets.*

**4.2 System Details**

After preprocessing the corpora with standard tok- enization tools, word-to-word alignments are per- formed in both directions, source-to-target and target-to-source. In our system implementation, the *GIZA++* toolkit3 is used to compute the word alignments. Then, the *grow-diag-final-and* (Koehn et al., 2005) heuristic is used to obtain the alignments from which tuples are extracted.

In addition to the tuple *n*-gram translation model, our SMT system implements six addi- tional feature functions which are linearly com-

1 [www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger](http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger)

2 [www.ims.uni-stuttgart.de/projekte/corplex/RFTagger](http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger)

3 [http://www.fjoch.com/GIZA++.html](http://www.fjoch.com/GIZA%2B%2B.html)

bined following a discriminative modeling frame- work (Och and Ney, 2002): a *target-language model* which provides information about the tar- get language structure and fluency; two *lexicon models*, which constitute complementary trans- lation models computed for each given tuple; a ’weak’ distance-based *distortion model*; and finally a *word-bonus model* and a *tuple-bonus model* which are used in order to compensate for the system preference for short translations.

All language models used in this work are estimated using the *SRI language modeling* toolkit4. According to our experience, Kneser- Ney smoothing (Kneser and Ney, 1995) and in- terpolation of lower and higher *n*-grams options are used as they typically achieve the best per- formance. Optimization work is carried out by means of the widely used *MERT* toolkit5 which has been slightly modified to perform optimiza- tions embedding our decoder. The *BLEU* (Pap- ineni et al., 2002) score is used as objective func- tion for MERT and to evaluate test performance.

**4.3 Reordering in German-English and**

**French-English Translation**

Two factors are found to greatly impact the overall translation performance: the morphological mis- match between languages, and their reordering needs. The vocabulary size is strongly influenced by the number of word forms for number, case, tense, mood, *etc.*, while reordering needs refer to the difference in their syntactic structure. In this work, we are primarily interested on the reorder- ing needs of each language pair. Figure 3 displays a quantitative analysis of the reordering needs for the language pairs under study.

Figure 3 displays the (%) distribution of the reordered sequences, according to their size, ob- served for the training bi-texts of both translation tasks. Word alignments are used to determine re- orderings. A reordering sequence can also be seen as the sequence of words implied in a reorder- ing rule. Hence, we used the reordering rules ex- tracted from the training corpus to account for re- ordering sequences. Coming back to the example of Figure 1, a single reordering sequence is found,

4 <http://www.speech.sri.com/projects/srilm/>

5 <http://www.statmt.org/moses/>

which considers the source words *perfect transla- tions*.

50

fr-en

de-en

45

40

35

30

25

20

15

10

5

2 3 4 5 6 7 >=8

Size (words)

Figure 3: *Size (in words) of reorderings (%) ob- served in training bi-texts.*

As can be seen, the French-English and German-English pairs follow a different distribu- tion of reorderings according to their size. A lower number of short-range reorderings are ob- served for the German-English task while a higher number of long-range reorderings. Considering mid-range reorderings (from 5 to 7 words), the

French-English pair shows a lower percentage (*∼*

14%) than the German-English (*∼* 22%). A simi-

lar performance is expected when considering the

opposite translation directions. Note that reorder- ings are extracted from word-alignments, an au- tomatic process which is far notoriously error- prone. The above statistics must be accordingly considered.

**4.4 Results**

Translation accuracy (BLEU) results are given in table 2 for the same *baseline* system performing different reordering models: source 6-gram LM (**sLM**); lexicalized reordering (**lex**); bilingual 6- gram LM (**bLM**) assuming a single POS-tagged version of each tuple. In the case of the German- English translation task we also report results for the bilingual 5-gram LM built from POS tags ob- tained from *RFTagger* containing a richer vocab- ulary tag set (**b**+**LM**). For comparison purposes, we also show the scores obtained by the **Moses** phrase-based system performing lexicalized re- ordering. Models of both systems are built sharing the same training data and word alignments.

The worst results are obtained by the **sLM** model. The fact that it only considers source- language information results clearly relevant to accurately model reordering. A very similar performance is shown by our bilingual *n*-gram system and Moses under lexicalized reordering (**bLM** and **Moses**), slightly lower results are obtained by the *n*-gram system under French- English translation.

test set. Table 3 shows the BLEU results of the reordering task. Bigram, trigram and 4gram pre- cision scores are also given.

Table 3: *Reordering accuracy (BLEU) results.*

|  |  |  |
| --- | --- | --- |
| *Pair* | *Config* | *BLEU (2g/3g/4g)* |
| *Fr❀En* | *lex bLM* | 71*.*69 (75*.*0*/*63*.*4*/*55*.*6)  71*.*98 (75*.*3*/*63*.*7*/*56*.*0) |
| *En❀Fr* | *lex*  *bLM* | 72*.*92 (75*.*5*/*65*.*0*/*57*.*6)  73*.*25 (75*.*8*/*65*.*4*/*58*.*1) |
| *De❀En* | *lex b*+*LM* | 62*.*12 (67*.*3*/*52*.*1*/*42*.*5)  63*.*29 (68*.*3*/*53*.*5*/*44*.*0) |
| *En❀De* | *lex*  *b*+*LM* | 62*.*72 (67*.*9*/*52*.*8*/*43*.*1)  63*.*36 (68*.*6*/*53*.*6*/*43*.*8) |

|  |  |  |
| --- | --- | --- |
| *Config* | *Fr❀En En❀Fr* | *De❀En En❀De* |
| *sLM lex bLM b*+*LM* | 22*.*32 21*.*97  22*.*46 22*.*09  **23***.***03 22***.***32**  *− −* | 17*.*11 12*.*23  17*.*31 12*.*38  17*.*37 12*.*58  **17***.***57 12***.***92** |
| *Moses* | 22*.*81 **22***.***33** | 17*.*22 12*.*45 |

Table 2: *Translation accuracy (BLEU) results.*

When moving from **lex** to **bLM**, our system increases its accuracy results for both tasks and translation directions. In this case, results are slightly higher than those obtained by Moses (same results for English-to-French). Finally, re- sults for translations performed with the bilingual *n*-gram reordering model built from rich German POS tags (**b**+**LM**) achieve the highest accuracy results for both directions of the German-English task. Even though results are consistent for all translation tasks and directions they fall within

the statistical confidence margin. Add *±*2*.*36

to French-English results and *±*1*.*25 to German-

English results for a 95% confidence level. Very

similar results were obtained when estimating our model for orders from 5 to 7.

In order to better understand the impact of the proposed reordering model, we have measured the accuracy of the reordering task. Hence, isolat- ing the reordering problem from the more general translation problem. We use BLEU to account the *n*-gram matching between the sequence of source words aligned to the 1-best translation hypothe- sis, *i.e.* the permutation of the source words out- put by the decoder, and the permutation of source words that monotonizes the word alignments with respect to the target reference. Note that in or- der to obtain the word alignments of the test sets we re-aligned the entire corpus after including the

As can be seen, the bilingual *n*-gram reordering model shows higher results for both translation tasks and directions than lexicalized reordering, specially for German-English translation. Our model also obtains higher values of *n*-gram pre- cision for all values of *n*.

Next, we validate the introduction of additional bilingual context in the translation process. Fig- ure 4 shows the average size of the translation unit *n*-grams used for the test set according to dif- ferent models (German-English), the surface form

3-gram language model (main translation model), and the new reordering model when built from the reduced POS tagset (POS) and using the rich POS tagset (POS+).

40 word-based bilingual units

POS-based bilingual units

POS+-based bilingual units

35

30

25

20

15

10

5

0

0 1 2 3 4 5 6 7 8 9

Size (units)

Figure 4: *Size of translation unit n-grams (%)*

*seen in test for different n-gram models.*

As expected, translation units built from the re- duced POS tagset are less sparse, enabling us to

introduce larger *n*-grams in the translation pro- cess. However, the fact that they achieve lower translation accuracy scores (see Table 2) indicates that the probabilities associated to these large *n*- grams are less accurate. It can also be seen that the model built from the rich POS tagset uses a higher number of large *n*-grams than the language model built from surface forms.

The availability of mid-range *n*-grams validates the introduction of additional bilingual context achieved by the new model, leading to effec- tively modeling mid-range reorderings. Notice additionally that considering the language model built from surface forms, only a few 4-grams of the test set are seen in the training set, which explains the small reduction in performance ob- served when translating with a bilingual 4-gram language model (internal results). Similarly, the results shown in Figure 4 validates the choice of using bilingual 5-grams for *b*+*LM* and 6-grams for *bLM* .

Finally, we evaluate the mismatch between the reorderings collected on the training data, and those output by the decoder. Table 4 shows the percentage of reordered sequences found for the

1-best translation hypothesis of the test set ac-

cording to their size. The French-to-English and

German-to-English tasks are considered.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pair | Config | 2 3 4 | 5 6 7 | *≥* 8 |
| *Fr❀En* | *lex bLM* | 58 23 10  57 23 11 | 5 2 1  4 2*.*5 1*.*5 | 1  1 |
| *De❀En* | *lex b*+*LM* | 33 24 22  35 25 19 | 14 5 1*.*5  13 5 2*.*5 | 0*.*5  0*.*5 |

Table 4: *Size (*%*) of the reordered sequences ob- served when translating the test set.*

Very similar distributions are observed for both reordering models. In parallel, distributions are also comparable to those presented in Figure 3 for reorderings collected from the training bi-text, with the exception of long-range and very short- range reorderings. This may be explained by the fact that system models, in special the distortion penalty model, typically prefer monotonic trans- lations, while the system lacks a model to support large-range reorderings.

**5 Conclusions and Further Work**

We have presented a new reordering model based on bilingual *n*-grams with units built from lin- guistic information, aiming at modeling the struc- tural adequacy of translations. We compared our new reordering model to the widely used lexical- ized reordering model when implemented in our bilingual *n*-gram system as well as using *Moses*, a state-of-the-art phrase-based SMT system.

Our model obtained slightly higher transla- tion accuracy (BLEU) results. We also analysed the quality of the reorderings output by our sys- tem when performing the new reordering model, which also outperformed the quality of those out- put by the system performing lexicalized reorder- ing. The back-off procedure used by standard language models allows to dynamically adapt the scope of the context used. Therefore, in the case of our reordering model, back-off allows to con- sider always as much bilingual context (*n*-grams) as possible. The new model was straightfor- ward implemented in our bilingual *n*-gram sys- tem by extending the log-linear combination im- plemented by our decoder. No additional decod- ing complexity was introduced in the baseline de- coding implementation.

Finally, we showed that mid-range reorder- ings are present in French-English and German- English translations and that our reordering model effectively tackles such reorderings. However, we saw that long-range reorderings, also present in these tasks, are yet to be addressed.

We plan to further investigate the use of differ- ent structural information, such as supertags, and tags conveying different levels of morphology in- formation (gender, number, tense, mood, *etc.*) for different language pairs.

**Acknowledgments**

This work has been partially funded by OSEO un- der the Quaero program.

**References**

F. Xia and M. McCord. Improving a Statistical MT System with Automatically Learned Rewrite Pat- terns. In *Proc. of the COLING 2004*, 508–514, Geneva, Switzerland, August 2004.

D. Chiang. Hierarchical phrase-based translation.

*Computational Linguistics*, 33(2):201–228, June

2007.

H. Hoang and Ph. Koehn. Improving Mid-Range Re- Ordering Using Templates of Factors. In *Proc. of the EACL 2009*, 372–379, Athens, Greece, March

2009.

J. M. Crego and J. B. Marin˜ o. Improving statistical MT by coupling reordering and decoding. In *Ma- chine Translation*, 20(3):199—215, July 2007.

Marin˜ o, Jose´ and Banchs, Rafael E. and Crego, Josep Maria and de Gispert, Adria and Lambert, Patrick and Fonollosa, J.A.R. and Costa-jussa`, Marta N- gram Based Machine Translation. In *Computa- tional Linguistics*, 32(4):527–549, 2006

Ch. Tillman. A Unigram Orientation Model for Sta- tistical Machine Translation. In *Proc. of the HLT- NAACL 2004*, 101–104, Boston, MA, USA, May

2004.

M. Collins, Ph. Koehn and I. Kucerova. Clause Re- structuring for Statistical Machine Translation. In Proc. of the ACL 2005, 531–540, Ann Arbor, MI, USA, June 2005.

Ph. Koehn, H. Hoang, A. Birch, Ch. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, Ch. Moran, R. Zens, Ch. Dyer, O. Bojar, A. Constantin and E. Herbst. Moses: Open Source Toolkit for Sta- tistical Machine Translation. In Proc. of the ACL

2007, demonstration session, prague, Czech Repub- lic, June 2007.

Y. Zhang, R. Zens and H. Ney Improved Chunk-level Reordering for Statistical Machine Translation. In Proc. of the IWSLT 2007, 21–28, Trento, Italy, Oc- tober 2007.

H. Schmid and F. Laws. Estimation of Conditional Probabilities with Decision Trees and an Applica- tion to Fine-Grained POS Tagging. In Proc. of the COLING 2008, 777–784, Manchester, UK, August

2008.

F.J. Och and H. Ney. Improved statistical alignment models. In *Proc. of the ACL 2000*, 440–447, Hong Kong, China, October 2000.

Ph. Koehn, A. Axelrod, A. Birch, Ch. Callison-Burch, M. Osborne and D. Talbot. Edinburgh System De- scription for the 2005 IWSLT Speech Translation Evaluation. In *Proc of the IWSLT 2005*, Pittsburgh, PA, October 2005.

F. J. Och and H. Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the ACL 2002*. 295–302, Philadelphia, PA, July 2002.

A. Stolcke. SRLIM: an extensible language model- ing toolkit. *Proc. of the INTERSPEECH 2002*. 901–

904, Denver, CO, September 2008.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. Bleu: a method for automatic evaluation of machine trans- lation. In *Proc. of the ACL 2002*, 311–318, Philadel- phia, PA, July 2002.

R. Kneser and H. Ney. Improved backing-off for m- gram language modeling. In *Proc. of the ICASSP*

*1995*. 181–184, Detroit, MI, May 1995.

A. Zollmann, A. Venugopal, F. J. Och and J. Ponte.

A Systematic Comparison of Phrase-Based, Hierar- chical and Syntax-Augmented Statistical MT. In *Proc. of the COLING 2008*. 1145–1152, Manch- ester, UK, August 2008.

M. Popovic and H. Ney. POS-based Word Reorderings for Statistical Machine Translation. In *Proc. of the LREC 2006*. 1278–1283, Genoa, Italy, May 2006.

J. Niehues and M. Kolss. A POS-Based Model for Long-Range Reorderings in SMT. In *Proc. of the WMT 2009*. 206–214, Athens Greece, March 2009.