Rule-based Reordering Space in Statistical Machine Translation

Nicolas Pécheux, Alexandre Allauzen, François Yvon

LIMSI/CNRS, B.P. 133, 91403 Orsay, France
Université Paris-Sud, 91403 Orsay, France
{pecheux,allauzen,yvon}@limsi.fr

Abstract

In Statistical Machine Translation (SMT), the constraints on word reorderings have a great impact on the set of potential translations that are explored. Notwithstanding computationnal issues, the reordering space of a SMT system needs to be designed with great care: if a larger search space is likely to yield better translations, it may also lead to more decoding errors, because of the added ambiguity and the interaction with the pruning strategy. In this paper, we study this trade-off using a state-of-the art translation system, where all reorderings are represented in a word lattice prior to decoding. This allows us to directly explore and compare different reordering spaces. We study in detail a rule-based preordering system, varying the length or number of rules, the tagset used, as well as contrasting with oracle settings and purely combinatorial subsets of permutations. We focus on two language pairs: English-French, a close language pair and English-German, known to be a more challenging reordering pair.

Keywords: Reordering Contraints, Empirical Analysis, Statistical Machine Translation

1. Introduction

Reordering is still a critical issue for statistical machine translation, and the reordering complexity for a language pair can be considered as a relevant indicator of the difficulty to automatically translate from one into the other (Birch et al., 2008). Reordering is problematic since the factorial space of permutations cannot be fully explored. Moreover, even if we could, this space would contain too much ambiguity and permutations that are linguistically meaningless. Therefore we must rely on methods that can restrict the space of possible reorderings. Various constraints on admissible permutations have been proposed in the past including IBM (Berger et al., 1996), MJ (Kumar and Byrne, 2005) or ITG (Wu, 1997). Those constraints have been compared in terms of performance (Zens and Ney, 2003; Zens et al., 2004) or in oracle settings (Dreyer et al., 2007; Wisniewski and Yvon, 2013). Other approaches include linguistically motivated rules that are automatically learned (Crego and Mariño, 2006; Niehues and Kolss, 2009; Popovic and Ney, 2006). To the best of our knowledge, these two families of approaches have not been compared yet.

In the phrase-based approach, sentences are first segmented into variable length segments or phrases, then reordered. Word reorderings can be divided in two tighly intertwined parts: local reorderings that take place within phrases; and longer reorderings of those phrases. Moreover, a recent trend has been to consider preordering methods, where source sentences are reordered in a pre-processing step to match the target word order and then fed into the standard Phrase-Based pipeline (Xia and McCord, 2004; Collins et al., 2005; Tromble and Eisner, 2009). This further complexifies the analysis of the word reorderings that are actually considered in translation. Finally, because of the pruning strategy, only a restricted part of the search space is effectively explored.

In this paper, we use a state-of-the-art n-gram SMT system (Crego et al., 2011), described in Section 2., that splits reordering and decoding into two separate steps. Reorderings of the source sentence are compactly encoded in a permutation lattice, the *reordering space*, that is then translated in a monotonic fashion. This allows us to study the reordering space that is explored and then to assess its impact on the whole translation process. Indeed, in addition to computationnal issues, there is a tradeoff when building the reordering space of a machine translation system. On the one hand, a larger space is more likely to contain a permutation that can yield a relevant translation. On the other hand, it may also cause more decoding errors, because of both the ambiguity of natural languages and the necessary pruning of the search space.

The main contribution of this work is to evaluate the impact of the reordering space on translation performance by exploring various experimental conditions (Section 3.). We study different methods to generate the reordering space by varying the word classes that are used by the reordering rules. Evaluation is carried out on two language pairs (French-English and German-English in both directions) that greatly differ by the range of the involved reorderings. The results in Section 4. show that our SMT system is not able to fully benefit from an accurate reordering space.

2. The n-gram Based Approach for SMT

For all our experiments, we use NCODE, an open source n-gram SMT toolkit[1], which achieved state-of-the-art performance in recent WMT evaluations (Callison-Burch et al., 2012; Bojar et al., 2013). NCODE implements the bilingual n-gram approach to SMT (Casacuberta and Vidal, 2004; Mariño et al., 2006; Crego and Mariño, 2006) that is closely related to the standard phrase-based approach. However, in this framework, the translation is divided into two steps: a source reordering step and a (monotonic) translation step. Since the translation step is monotonic, the translation model relies on the n-gram assumption to de-

[1] http://ncode.limsi.fr

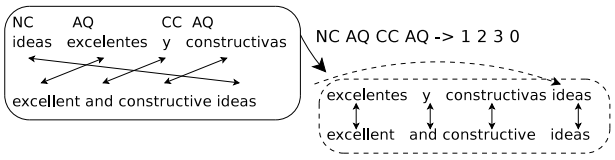

Figure 1: *Reordering rules extraction from word alignment. This figure is borowed from (Crego and Mariño, 2006).*

compose the joint probability of a sentence pair in a sequence of bilingual units called *tuples*.

In addition to the translation model, NCODE uses a set of feature functions embedded in a log-linear model (Och and Ney, 2002) that is similar to standard phrase-based systems (see Crego et al. (2011) for details). The models that have an impact on the selected reordering are the monolingual and bilingual n-gram models, the *lexicalized reordering models* (Tillmann, 2004; Crego et al., 2011) that aim at predicting the orientation of the next translation unit and a "weak" distance-based *distortion model*.

During training, source sentences are first reordered so as to match the target word order by unfolding the word alignments, producing here called *unfolded reorderings* (see figure 1). Tuples are then extracted in such a way that a unique segmentation of the bilingual corpus is achieved. A n-gram translation model is finally estimated over the training corpus composed of tuple sequences using modified Knesser-Ney Smoothing (Chen and Goodman, 1998).

During decoding, a source sentence is represented in the form of a word lattice, so as to reproduce a set of possible word order modifications introduced during the tuple extraction process. This lattice represents the reordering space that is then searched for the best possible candidate translation. As exhaustive search is intractable, NCODE uses a beam search strategy based on stacks. As futur cost estimation is problematic for multiple n-gram models, NCODE uses one stack per hypothesis translating the *same input words*, in contrast to the *same number of words* as in standart Phrase-Based systems. Thus the complexity depends on the number of nodes in the reordering lattice.

3. Generation of Reordering Lattices

In NCODE source reordering is based on a set of rewriting rules that non-deterministically reorder the input words. In this section, we explain in more details the reordering lattice mechanism used in NCODE, as well as variants considered in our experiments.

3.1. Reordering Rules Extraction

Reordering rules are automatically learned during the unfolding procedure as depicted in Figure 1. Let $\mathbf{w} = w_1 w_2 \ldots w_n$ be a source sentence and $\mathbf{t} = t_1 t_2 \ldots t_n$ an associated tags sequence. Let $\mathbf{w}_\sigma = w_{\sigma_1} w_{\sigma_2 \ldots} w_{\sigma_n}$ be the reordered sentence obtained by the unfolding procedure where $\sigma = \sigma_1 \ldots \sigma_n$ is a permutation, $\sigma \in \mathfrak{S}_n$ the set of permutations of $\{1, \ldots, n\}$. A reordering rule is extracted for any subsequence $\sigma_{[i:j]} = \sigma_i \ldots \sigma_j$ of σ with $|j - i| \geq 1$ such as

$$i \leq k \leq j \;\Rightarrow\; i \leq \sigma_k \leq j,$$

and that is minimal under this property. Rules have the following form:

$$\mathbf{t}_{[i:j]} \rightarrow \bar{\sigma}_{[i:j]}$$

where $\bar{\sigma}_{[i:j]}$ is the induced permutation in $\mathfrak{S}_{|j-i+1|}$ obtained by renumbering $\sigma_{[i:j]}$. Spans $\mathbf{w}_{[i:j]}$ correspond to the smallest (non trivial) ones to be reordered in order to recover $\mathbf{w}$. One could also extract all rules $\mathbf{t}_{[i:j]} \rightarrow \bar{\sigma}_{[i:j]}$ for any span $|j - i| \geq 1$, but previous experiments showed a slight drop in performance.

To filter out alignment noise and to limit the size of the reordering space, rules may be pruned according to a maximum cost threshold (by default 4):

$$\text{cost}(\mathbf{t} \rightarrow \sigma) = -\log \frac{\text{count}(\mathbf{t} \rightarrow \sigma)}{\sum_{\sigma' \in \mathfrak{S}_{|\mathbf{t}|}} \text{count}(\mathbf{t} \rightarrow \sigma')}$$

where $\mathbf{t}$ is any tag subsequence, $\sigma \in \mathfrak{S}_{|\mathbf{t}|}$ a permution and the counts are computed on the training data. Since this cost is the negative logarithm of a conditional ratio, a coarser tagset might be more heavily pruned, resulting in a smaller set of extracted rules.

Rules may be also pruned according to their length (by default 10). Previous experiments show that further increasing this length hardly makes any difference. In fact, long rules are too sparse to possibly generalize beyond the training set. Long range reorderings are thus explicitly excluded from the model.

3.2. Reordering Lattices Generation

For any sentence $\mathbf{w}$ with tags $\mathbf{t}$ we start with a lattice containing the monotonic path $\mathbf{w}$. For each segment $\mathbf{w}_{[i:j]}$ and each rule $\mathbf{t}_{[i:j]} \rightarrow \sigma$ we add the subpath $\sigma(\mathbf{w}_{[i:j]})$ to the lattice. Note this is done in a parallel fashion so that rewriting rules do not interfere with each other. Applying the reording rules finally results in a finite-state graph that represents the *reordering space*.

3.3. Alternative Tagsets

Rewriting rules are built using Part-of-speech (POS), rather than surface word forms (Crego and Mariño, 2006) to increase their generalization power. However, any word factor may be possibly used. To investigate different levels of generalization and the relevance of syntactic word factors, different tagsets are introduced.

- **Simplified POS** (spos): The tagset is reduced to 12 simple language-independent categories, in an attempt to limit the sparsity of the extracted rules. This tagset has been designed independenlty, but turn out to be very close to the universal POS tagset described in (Petrov et al., 2012). For under resourced languages, universal POS can be projected by cross-lingual transfer or learned from weak annotations (Li et al., 2012; Täckström et al., 2013), thereby relaxing the need for a POS tagger.

- **Enhanced POS** (e50pos): The POS tags are lexicalized for the 50 most frequent words, resulting in more specific rules. Enhanced tags are closely related to lexicalized rules (Huang and Pendus, 2013).

```
(s) : the meeting was announced by the president's spokesman Radim Ochvat .
(r) : c' est le porte-parole présidentiel Radim Ochvat qui a informé de la réunion .
(m) : la réunion a été annoncée par le président porte-parole Radim Ochvat .
(l) : la réunion a été annoncée par le porte-parole du président Radim Ochvat .
(u) : le porte-parole du président Radim a été annoncée par la réunion Ochvat .
(o) : de la réunion a informé de la c' est le porte-parole présidentiel Radim Ochvat qui
```

Figure 2: *Translations of a source sentence* `(s)` *from* newstest2010*, along with the reference translation* `(r)`*, contrasting monotone* `(m)` *lattice based* `(l)` *and unfolded reordering* `(u)` *constraints, as well as oracle decoding* `(o)` *in the lattice reordering space.*

- **Brown classes** (classes): Statistical word classes were found to be a good approximation for Part-of-Speech tags when a POS tagger is not available. In (Ramanathan and Visweswariah, 2012), word clusters perform worse than POS, but still reasonably well, in a preordering setting. In this work, we compute the statistical word classes using the methods of Brown et al. (1992).

3.4. MJ-i

In principle, one can design any kind of permutation constraints and encode them in a lattice. In practice, the number of nodes in the lattice must remain reasonable (polynomial) in the number of words in the sentence.
To assess whether constraining the reorderings to those observed in the data is appropriate, we contrast rule-based approaches with MaxJump (MJ) constraints (Kumar and Byrne, 2005). In MJ-i, a word move cannot exceed i positions. This is equivalent to using a rule-based system, where all possible rules up to size $i + 1$ are considered.

3.5. Metrics and Unfolded Reorderings

Given our assumptions, the reordering space should contain the unfolded reordering as defined in Section 2. For unseen data, this oracle can be derived from forced alignments between source sentences and their associated references. Therefore, as a quality measure on reordering constraints, we define the *coverage* on some test set as the number of time the reordering space contains the reference reordering. We also compute the *size* of the reordering space as the number of paths[2] and edges in the reordering lattice.[3]
In that sense, the best reordering constraints should be the ones that generate lattices containing the unfolded reordering as only alternative. We refer to this oracle-like constraint as *unfolded reordering* constraint.

4. Experimental Results

4.1. Data and System Description

We considered two different tasks: 1) the French-English Basic Traveling Expression Corpus (BTEC) (Paul et al., 2010), using *train10*, *devel03* and *test09* for training, tuning and testing, respectively; 2) The English-French and English-German training data of NEWSCOMMENTARY provided by organizers of WMT'12 (Callison-Burch et al., 2012), with *newstest2009* and *newstest2010* for tuning and for testing, respectively.
All data is preprocessed as described in (Allauzen et al., 2013). For each task, a 4-gram language model is estimated using the target side of the training data and 50 word classes[4] derived using MKCLS.[5] We used NCODE with the default setting and an additional a POS-POS bilingual factor model.[6] Beam size was set to 25 during the tuning step and to 50 when decoding, as this showed some gains in previous experiments. Oracles are computed using lattice minumum Bayes-risk decoding with linear corpus BLEU as described in Tromble et al. (2008) using unigram precision and recall values of 0.8. All results are averaged over 5 runs of MERT to control for optimizer instability (Clark et al., 2011). Approximate randomization tests for multiple optimizer samples to assess statistical significance are carried out using MULTEVAL.[7]

4.2. From monotone to Oracle Reorderings

Table 1 shows BLEU scores on test data for the BTEC and NEWSCOMMENTARY task, for three different reordering spaces of increasing "quality". The first reordering space only considers the original source sentence order (monotone). The second uses our rule-based approach to create the reordering lattice. The last and most specific one considers exactly the reordering that match the target order (forced unfolded reordering). An example of sentence translation for those configurations is shown in Figure 2. Monotone translation does not succeed in inverting `president's` and `spokesman`, thus resulting in a mistranslation (meaning "by the president, spokesman, Radim Ochvat").
We can observe BLEU improvement from monotone to lattice reordering, as one would expect. For French-English, the increase may be as high as 3 BLEU points. However for English-German, the gain is much lower, especially for $en \rightarrow de$ direction (only about a half BLEU point). This could suggest that our reordering system does not succeed in predicting German word order.
Although direct comparison of BLEU scores on different corpora is unfair, scores for $en \rightarrow de$ are always worst, indicating a more challenging language direction. German is a morphologically rich language that exhibits long range

[2]Computed efficiently using the counting semiring.

[3]The number of edges closely relates to the decoding complexity.

[4]Out-of-vocabulary are mapped to class 1.

[5]`http://code.google.com/p/giza-pp/`

[6]Note that this is independent of the choice of tags used in the reordering rules.

[7]`https://github.com/jhclark/multeval`

			BTEC task		NEWSCOMMENTARY task			
	tun.	dec.	en → fr	fr → en	en → fr	fr → en	en → de	de → en
ncode	(m)	(m)	43.4$_{\pm 0.4}$	46.8$_{\pm 0.4}$	20.3$_{\pm 0.1}$	20.9$_{\pm 0.1}$	12.7$_{\pm 0.0}$	17.4$_{\pm 0.1}$
	(l)	(l)	46.8$_{\pm 0.2}$	49.1$_{\pm 0.4}$	23.3$_{\pm 0.1}$	23.0$_{\pm 0.1}$	13.2$_{\pm 0.1}$	18.5$_{\pm 0.1}$
	(u)	(u)	48.7$_{\pm 0.3}$	51.6$_{\pm 0.6}$	25.4$_{\pm 0.1}$	27.0$_{\pm 0.1}$	15.9$_{\pm 0.0}$	22.3$_{\pm 0.0}$
oracle		(m)	88.6	86.4	70.3	70.5	55.7	64.0
		(l)	96.4	95.2	84.3	84.3	64.6	74.0
		(u)	98.8	99.7	92.7	94.7	81.9	92.7
ncode	(u)	(l)	43.4$_{\pm 5.6}$	45.8$_{\pm 2.7}$	23.1$_{\pm 0.3}$	21.7$_{\pm 1.5}$	13.0$_{\pm 0.0}$	18.1$_{\pm 0.2}$
	(l)	(u)	48.7$_{\pm 0.2}$	52.2$_{\pm 0.5}$	25.4$_{\pm 0.1}$	26.9$_{\pm 0.1}$	15.9$_{\pm 0.0}$	22.3$_{\pm 0.1}$

Table 1: *BLEU scores on test data obtained by* NCODE *system and oracle decoding, when no reorderings are allowed (monotone (m)), using our lattice reordering space (l) and when given exactly the unfolded reordering (u), during tuning phase on development data (tun.) and when decoding the test (dec.). Reported BLEU scores are averages across 5 runs of MERT along with standart deviation across runs shown in script size.*

	maxcost	BLEU ncode	BLEU oracle	#rules	size	coverage (%)
$en \rightarrow fr$	0	19.6*$_{\pm 0.1}$	70.3	0	27 / 1	19
	2	22.6*$_{\pm 0.1}$	81.3	20k	34 / 40	40
	4	22.8$_{\pm 0.1}$	84.6	30k	49 / 10^4	50
	8	22.3*$_{\pm 0.2}$	87.9	42k	217 / 10^{21}	62
$fr \rightarrow en$	0	20.2*$_{\pm 0.1}$	70.5	0	30 / 1	16
	2	21.9*$_{\pm 0.0}$	77.2	19k	35 / 11	25
	4	22.5$_{\pm 0.1}$	84.4	31k	69 / 10^7	40
	8	22.2*$_{\pm 0.1}$	89.6	49k	397 / 10^{34}	60
$en \rightarrow de$	0	12.7*$_{\pm 0.1}$	55.7	0	27 / 1	16
	2	12.7*$_{\pm 0.1}$	57.0	61k	30 / 2.1	17
	4	13.1$_{\pm 0.1}$	64.6	83k	65 / 10^4	25
	8	12.9*$_{\pm 0.1}$	70.0	99k	250 / 10^{21}	33
$de \rightarrow en$	0	17.6*$_{\pm 0.1}$	64.1	0	28 / 1	15
	2	17.9*$_{\pm 0.1}$	66.3	63k	33 / 3.5	17
	4	18.7$_{\pm 0.1}$	74.1	86k	67 / 10^6	25
	8	18.7$_{\pm 0.1}$	79.8	103k	264 / 10^{34}	33

Table 2: *Impact of rule filtering strategy (maxcost) in* NEWSCOMMENTARY *task on: BLEU scores obtained by* NCODE *system and oracle decoding; the number of reordering rules (#rules); the size of the lattice reordering space (averaged number of arcs / average number of paths); and on the coverage (see section 3.5.).* Reported BLEU scores are averages across 3 runs of MERT along with standart deviation ac ross runs shown in script size. A statistical significance ($p < 0.005$) difference from the $maxcost = 4$ baseline is indicated by * symbol.

reorderings when translating from English.

All language directions also benefit from being given the unfolded reordering. Note however that a better BLEU does not always result in a better translation. In the example in Figure 2, the lattice translation is perfectly valid although different from the reference. However, the translation in the unfolded reordering condition is mistranslated as the initial sentence is in the passive voice (the translated sentence means "President Radim's spokesman has been announced by the Ochvat meeting"). This example shows that artificial reorderings may sometimes lead to poor translations. Yet, an improvement up to 4 BLEU points when translating toward English shows that there is a room for improvement. However the improvement for $en \rightarrow de$ is not so clear. In other words, "solving" the reordering problem at decoding time has only a slight effect on performance for this language direction. Therefore the reordering constraints might currently not be the main limitation for our system. It is worth noticing that the reordering length is limitless in the unfolded reordering case, hence the lack of long range reorderings in our model can not be the main explanation.

Table 1 also shows best possible BLEU scores (oracle) for the three conditions. We can observe a positive correlation between oracle BLEU scores and the one obtained by the system. These high oracle BLEU scores also suggest that a larger gain may be expected from improving the translation models rather than by increasing the reordering space. It is worth noticing that oracle BLEU scores are highly op-

	$en \rightarrow fr$		$fr \rightarrow en$		$en \rightarrow de$		$de \rightarrow en$	
	BLEU	size	BLEU	size	BLEU	size	BLEU	size
maxlen=2	$22.4_{\pm 0.1}$	34 / 10^2	$21.9_{\pm 0.0}$	40 / 10^3	$12.8_{\pm 0.0}$	34 / 10^2	$17.6_{\pm 0.1}$	34 / 30
MJ-1	$22.1^*_{\pm 0.1}$	74 / 10^{14}	$21.9_{\pm 0.0}$	84 / 10^{17}	$12.8_{\pm 0.0}$	74 / 10^{14}	$17.8^*_{\pm 0.2}$	19 / 10^{19}
maxlen=3	$22.6_{\pm 0.0}$	40 / 10^3	$22.2_{\pm 0.0}$	47 / 10^5	$12.9_{\pm 0.1}$	42 / 10^3	$18.0_{\pm 0.1}$	41 / 10^3
MJ-2	$22.3^*_{\pm 0.1}$	209 / 10^{23}	$22.2_{\pm 0.1}$	239 / 10^{28}	$12.9_{\pm 0.1}$	209 / 10^{23}	$18.0_{\pm 0.1}$	223 / 10^{31}
maxlen=4	$22.9_{\pm 0.2}$	43 / 10^4	$22.3_{\pm 0.1}$	56 / 10^6	$13.0_{\pm 0.1}$	50 / 10^4	$18.3_{\pm 0.1}$	50 / 10^4
MJ-3	$22.4^*_{\pm 0.1}$	715 / 10^{30}	$22.1^*_{\pm 0.1}$	824 / 10^{36}	$12.8^*_{\pm 0.0}$	715 / 10^{30}	$17.8^*_{\pm 0.2}$	768 / 10^{40}

Table 3: *Comparison between rule-based reordering with a length rule limit (maxlen) and purely combinatorial MaxJump constraints (MJ-i) for* NEWSCOMMENTARY *task. Reported BLEU scores are averages across 3 runs of MERT along with standart deviation across runs shown in script size. A statistical significance ($p < 0.005$) difference between $maxlen = i$ and MJ-$(i-1)$ is indicated by * symbol.*

timistic and a large part of the gain may result from overfitting the BLEU metric. An illustration is provided in Figure 2 with the mumbo-jumbo oracle translation.

We also explored the importance of the reordering space during the tuning step. Table 1 shows that when tuning with unfolded reorderings instead of a lattice, we observe a small drop for all directions when decoding on a lattice along with a dramatic increase in the optimizer variance. This means that the reordering features do benefit to see the whole reordering space during tuning.

4.3. Reordering Space Trade-off

Table 2 shows reordering space size, coverage, oracle and decoding scores for various constraints (results for the BTEC task are omitted for brievty). We observe that while the number of rules is almost twice for $en \rightarrow de$ than for $en \rightarrow fr$, the generated reordering spaces are comparable in sizes, but with a much lower coverage for $en \rightarrow de$.

By relaxing the rules pruning, we see large increases in reordering space size, in coverage and in oracle BLEU. However, in regular test condition, we observe a degradation of the BLEU scores, when the size of the reordering space drastically increases. This shows the importance of the trade-off regarding the design of the reordering space.

4.4. Alternative Tagsets

	en $\rightarrow$ fr	fr $\rightarrow$ en	en $\rightarrow$ de	de $\rightarrow$ en
baseline	$22.8_{\pm 0.1}$	$22.5_{\pm 0.1}$	$13.1_{\pm 0.1}$	$18.7_{\pm 0.1}$
spos	$22.7_{\pm 0.1}$	$22.5_{\pm 0.2}$	$13.1_{\pm 0.0}$	$18.7_{\pm 0.1}$
e50pos	$22.8_{\pm 0.1}$	$22.5_{\pm 0.1}$	$13.1_{\pm 0.1}$	$18.3^*_{\pm 0.1}$
classes	$22.8_{\pm 0.1}$	$22.6_{\pm 0.1}$	$12.9^*_{\pm 0.1}$	$18.2^*_{\pm 0.2}$

Table 4: *Comparison of different tagsets for* NEWSCOMMENTARY *task. Reported BLEU scores are averages across 5 runs of MERT along with standart deviation across runs shown in script size. A statistical significance ($p < 0.005$) difference from the baseline is indicated by * symbol.*

From Table 4 we observe that the tagset has limited influence on BLEU scores. Oracle scores, spaces size, and coverage were also very similar. Moreover, the competitive results obtained with the coarse grained tagset and the automatic word classes show that they can be used as a workaround for under resourced language.

4.5. Comparison with MJ-i

Table 3 provides a head to head comparison between MJ-i constraints and the rule based approach. The MJ reordering spaces are several orders of magnitude larger than their ruled-based counterpart but yield to significanlty lower results. This justifies the use of linguistically motivated rules, instead of allowing all local permutations and corroborate the trade-off discussed earlier. Training time is also an issue here: for $en \rightarrow fr$, the tuning step with MJ-3 constraints takes four times as long than $maxlen = 4$.

5. Conclusions

In this work, we have compared the reordering space generated by different reordering rules as well as local permutation constraints. We use a n-gram SMT tool that separates reordering and decoding, but our approach is more general as soon as the reordering space may be encoded in a lattice prior to decoding. We compare the different reordering constraints from an oracle point of view, but also taking into account the trade-off between expressivity and size, exploring the interaction with a pruned decoding.

6. Acknowledgements

We would like to thank Thomas Lavergne and Guillaume Wisniewski for their helpful feedback and for having provided the oracle and semiring frameworks used in this work. We would also like to thank to our anonymous reviewers for their comments and suggestions.

7. References

Alexandre Allauzen, Nicolas Pécheux, Quoc Khanh Do, Marco Dinarelli, Thomas Lavergne, Aurélien Max, Hai-Son Le, and François Yvon. 2013. LIMSI @ WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 62–69, Sofia, Bulgaria.

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Andrew S. Kehler, and Robert L.

Mercer. 1996. Language translation apparatus and method using context-based translation models.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of EMNLP*, Honolulu, USA.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistic*, 18(4):467–479.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Francesco Casacuberta and Enrique Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(3):205–225.

Stanley F. Chen and Joshua T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation : Controlling for Optimizer Instability. In *Better Hypothesis Testing for Statistical Machine Translation : Controlling for Optimizer Instability*, pages 176–181, Portland, Oregon.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 531–540, Ann Arbor, Michigan.

Josep M. Crego and José B. Mariño. 2006. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

Josep M. Crego, François Yvon, and José B. Mariño. 2011. N-code: an open-source bilingual N-gram SMT toolkit. *Prague Bulletin of Mathematical Linguistics*, 96:49–58.

Markus Dreyer, Keith B. Hall, and Sanjeev P. Khudanpur. 2007. Comparing reordering constraints for smt using efficient BLEU oracle computation. In *NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 103–110, Rochester, New York.

Fei Huang and Cezar Pendus. 2013. Generalized Reordering Rules for Improved SMT. In *Association for Computer Linguistics*, volume 25380, pages 387–392, Sofia, Bulgaria. The Association for Computer Linguistics.

Shankar Kumar and William Byrne. 2005. Local phrase reordering models for statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 161–168, Vancouver, British Columbia, Canada.

Shen Li, João V Graça, and Ben Taskar. 2012. Wiki-ly Supervised Part-of-Speech Tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1389—-1398, Korea. Association for Computational Linguistics.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrick Lambert, José A.R. Fonollosa, and Marta R. Costa-Jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

Jan Niehues and Muntsin Kolss. 2009. A POS-based model for long-range reorderings in SMT. In *Fourth Workshop on Statistical Machine Translation*, Athens, Greece.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, Pennsylvania, USA.

Michael Paul, Marcello Federico, and Sebastian Stücker. 2010. Overview of the IWSLT 2010 Evaluation Campaign. In *Proc. IWSLT*, pages 3–27.

Slav Petrov, Dipanjan Das, and Ryan Mcdonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey. European Language Resources Association.

M. Popovic and H. Ney. 2006. Pos-based word reordering for statistical machine translation. In *LREC*.

Ananthakrishnan Ramanathan and Karthik Visweswariah. 2012. A Study of Word-Classing for MT Reordering. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 3971–3976, Istanbul, Turkey.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan Mcdonald, and Joakim Nivre. 2013. Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL*, pages 101–104.

Roy Tromble and Jason Eisner. 2009. Learning linear ordering problems for better translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1007–1016, Stroudsburg, PA, USA.

Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice Minimum Bayes-Risk decoding for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 620–629, Honolulu, Hawaii.

Guillaume Wisniewski and François Yvon. 2013. Oracle decoding as a new way to analyze phrase-based machine translation. *Machine Translation*, 28(2):1–24.

Dekai Wu. 1997. Stochastic inversion transduction gram-

mars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 508–514, Geneva, Switzerland.

Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *ACL*, pages 144–151.

Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *COLING*, pages 205–211, Geneva, Switzerland.