Using syntactic coupling features for discriminating phrase-based translations (WMT-08 Shared Translation Task)

Vassilina Nikoulina and Marc Dymetman
Xerox Research Centre Europe
Grenoble, France
{nikoulina,dymetman}@xrce.xerox.com

Abstract

Our participation in the shared translation task at WMT-08 focusses on news translation from English to French. Our main goal is to contrast a baseline version of the phrase-based MATRAX system, with a version that incorporates syntactic "coupling" features in order to discriminate translations produced by the baseline system. We report results comparing different feature combinations.

1 Introduction

Our goal is to try to improve the fluency and adequacy of a baseline phrase-based SMT system by using a variety of "syntactic coupling features", extracted from parses for the source and target strings. These features are used for reranking the n-best candidates of the baseline system.

The phrase-based SMT system MATRAX, developed at XRCE, is used as the baseline in the experiments. MATRAX is based on a fairly standard log-linear model, but one original aspect of the system is the use of non-contiguous bi-phrases such as *ne ... plus / not ... anymore*, where words in the source and target phrases may be separated by gaps, to be filled at translation time by lexical material provided by some other such pairs (Simard et al., 2005).

For parsing, we use the *Xerox Incremental Parser* XIP (Aït-Mokhtar et al., 2002), which is a robust dependency parser developed at the Xerox Research Centre Europe. XIP is fast (around 2000 words per second for English) and is well adapted to a situation, like the one we have here, were we need to parse on the order of a few hundred target candidates on the fly. Also of interest to us is the fact that XIP produces labelled dependencies, a feature that we use in some of our experiments.

1.1 Decoding and Training

We resort to a standard reranking approach in which we produce an n-best list of MATRAX candidate translations (with n = 100 in our experiments), and then rerank this list with a linear combination of our parse-dependent features. In order to train the feature weights, we use an averaged structured perceptron approach (Roark et al., 2004), where we try to learn weights such that the first candidate to emerge is equal to the "oracle" candidate, that is, the candidate that is closest to the reference in terms of NIST score.

1.2 Coupling Features

Our general approach to computing coupling features between the dependency structure of the source and that of a candidate translation produced by MATRAX is the following: we start by aligning the words between the source and the candidate translation, we parse both sides, and we count (possibly according to a weighting scheme) the number of configurations ("rectangles") that are of the following type: $((s_1, s_{12}, s_2), (t_1, t_{12}, t_2))$, where s_{12} is an edge between s_1 and s_2, t_{12} is an edge between t_1 and t_2, s_1 is aligned with t_1 and s_2 is aligned with t_2. We implemented several variants of this basic scheme.

We start by describing different "generic" coupling functions derived from the basic scheme, as-

Proceedings of the Third Workshop on Statistical Machine Translation, pages 159–162,
Columbus, June 2008.

suming that word alignments have been already determined, then we describe the option of taking into account specific dependency labels when counting rectangles, and finally we describe two options for computing the word alignments.

1.2.1 Generic features

The first measure of coupling is based on simple, non-weighted, word alignments. Here we simply consider that a word of the source and a word of the target are aligned or not aligned, without any intermediary degree, and consider that a rectangle exists on the quadruple of words s_1, s_2, t_1, t_2 iff s_i is aligned to t_i, s_1 and s_2 have a dependency link between them (in whatever direction) and similarly for t_1 and t_2. The first feature that we introduce, *Coupling-Count*, is simply the count of all such rectangles between the source and the target.

We note that the value of this feature tends to be correlated with the size of the source and target dependency trees. We therefore introduce some normalized variants of the feature:

- *Coupling-Recall*. We compute the number of source edges for which there exists a projection in the target. More formally, the number of edges between two words s_1, s_2 such that there exist two words t_1, t_2 with s_i aligned to t_i and such that t_1, t_2 have an edge between them. We then divide this number by the total number of edges in the source.

- *Coupling-Precision*. We do the same thing this time starting from the target.

- *Coupling-F-measure*. This is defined as the harmonic mean of the two previous features.

1.2.2 Label-specific features

The features previously defined do not take into account the labels associated with edges in the dependency trees. However, while rectangles of the form $((s_1, \text{subj}, s_2), (t_1, \text{subj}, t_2))$ may be rather systematic between such languages as English and French, other rectangles may be much less so, due on the one hand to actual linguistic divergences between the two languages, but also, as importantly in practice, to different representational conventions used by different grammar developers for the two languages.[1]

In order to control this problem, we introduce a collection of *Label-Specific-Coupling* features, each for a specific pair of source label and target label. The values of a label-specific feature are the number of occurrences for this specific label pair. We use only label pairs that have been observed to be aligned in the training corpus (that is, that participate in observed rectangles). In one version of that approach, we use all such pairs found in the corpus, in another version only the pairs above a certain frequency threshold in the corpus.

1.2.3 Alignment

In order to compute the features described above, a prerequisite is to be able to determine a word alignment between the source and a candidate translation. Our first approach is to use GIZA++ (corresponding roughly to IBM Model 4) to create these alignments, by producing for a given source and a given candidate translation n-best alignment lists in both directions and applying standard techniques of symmetrization to produce a bidirectional alignment.

Another way to find word alignments is to use the information provided by the baseline system. Since MATRAX is a phrase-based system, it has access to the bi-phrases (aligned by definition) that are used in order to generate a candidate translation. However note that when we use a bi-phrase based alignment, there will be differences from the word alignment that we discussed before, and we need to adapt our coupling functions.

1.2.4 Related approaches

There is a growing body of work on the use of syntax for improving the quality of SMT systems. Our approach is closest to the line taken in (Och et al., 2003), where syntactic features are also used for discriminating between candidates produced by a phrase-based system, but here we introduce and compare results for a wider variety of coupling features, taking into account different combinations involving normalization of the counts, symmetrized features between the source and target, labelled de-

[1] Although the XIP formalism is shared between grammar developers of French and English, the grammars do sometimes follow different conventions.

pendencies, and also consider several ways for computing the word alignment on the basis of which edge couplings are determined.

2 Experiments

2.1 Description

Our participation concerns the English to French News translation task. To train our baseline system we used the News Commentary corpus, namely the training ($\sim$ 1M words) and development (1057 sentences) sets proposed for the shared translation task. The same development set was used for the MERT training procedure of the baseline system, as well as for learning the parameters of the reranking procedure. Note that the test data on which we report our experimental results here is the one proposed as development test set for the News translation task (1064 sentences, nc-devtest2007).

Using MATRAX as the baseline system we generate 100-best lists of candidate translations for all source sentences of the test set, we rerank these candidates using our features, and we output the top candidate. We present our results in Table 1, distinguished according to the actual combination of features used in each experiment.

- The *Baseline* entry in the table corresponds to MATRAX results on the test set, without the use of any of the coupling features.

- We distinguish two sub-tables, according to whether Giza-based alignments or phrase-based alignments were used.

- The *Generic* keyword corresponds to the coupling features introduced in section 1.2.1, based on rectangle counts, independent of the labels of the edges.

- The *Matrax* keyword corresponds to using MATRAX "internal" features as reranking features, along with the coupling features. These MATRAX features are pretty standard phrase-based features, apart from some features dealing explicitly with gapped phrases, and are described in detail in (Simard et al., 2005).

- The *Labels* and *Frequent Labels* keywords corresponds to using label-specific features. In the first case (Labels) we extracted all of the aligned label pairs (label pair associated with a coupling rectangle) found in a training set, while in the second case (Frequent Labels), we only kept the most frequently observed among these label pairs.

- When several keywords appear on a line, we used the union of the corresponding features, and in the last line of the table, we show a combination involving at the same time some features computed on the basis of Giza-based alignments and of phrase-based alignments.

- Along with the NIST and BLEU scores of each combination, we also conducted an informal manual assessment of the quality of the results relative to the MATRAX baseline. We took a random sample of 100 source sentences from the test set and for each sentence, assessed whether the first candidate produced by reranking was better, worse, or indistinguishable in terms of quality relative to the baseline translation. We report the number of improvements (+) and deteriorations (-) among these 100 samples as well as their difference.[2]

3 Discussion

While the overall results in terms of Bleu and Nist do not show major improvements relative to the baseline, there are several interesting observations to make. First of all, if we focus on feature combinations in which MATRAX features are included (shown in italics in the table), we see that there is a general tendency for the results, both in terms of automatic and human evaluations, to be better than for the same combination without the MATRAX features; the explanation seems to be that if we do not use the MATRAX features during reranking, but consider the 100 candidates in the n-best list to be equally valuable from the viewpoint of MATRAX features, we lose essential information that cannot

[2] All the results reported here correspond to our own evaluations, prior to the WMT evaluations. Given time constraints, we focussed more on contrasting the baseline with the baseline + coupling features, than in tuning the baseline itself for the task at hand. After the submission deadline, we were able to improve the baseline for this task.

	NIST	BLEU	-	+	Diff
Baseline	6.4093	0.2034	0	0	0
Giza-based alignments					
Generic	6.3383	0.2043	15	17	2
Generic, Matrax	*6.3782*	*0.2083*	*4*	*18*	*14*
Labels	6.3483	0.1963	12	18	6
Labels, Generic	6.3514	0.2010	3	18	15
Labels, Generic, Matrax	*6.4016*	*0.2075*	*3*	*20*	*17*
Frequent Labels	6.3815	0.2054	7	11	4
Frequent Labels, Generic	6.3826	0.2044	6	18	12
Frequent Labels, Generic, Matrax	*6.4177*	*0.2100*	*2*	*16*	*14*
Phrase-based alignments					
Generic	6.2869	0.1964	12	14	2
Generic, Matrax	*6.3972*	*0.2031*	*4*	*11*	*7*
Labels	6.3677	0.1995	16	15	-1
Labels, Generic	6.3567	0.1977	8	15	7
Labels, Generic, Matrax	*6.4269*	*0.2049*	*4*	*17*	*13*
Frequent Labels	6.3701	0.1998	3	15	12
Frequent Labels, Generic	6.3846	0.2013	7	16	9
Frequent Labels, Generic, Matrax	*6.4160*	*0.2049*	*4*	*16*	*12*
Giza Generic, Phrase Generic, Giza Labels, Matrax	*6.4351*	*0.2060*	*7*	*22*	*15*

Table 1: Reranking results.

be recovered simply by appeal to the syntactic coupling features.

If we now concentrate on the lines which do include MATRAX features and compare their results with the baseline, we see a trend for these results to be better than the baseline, both in terms of automatic measures as (more strongly) in terms of human evaluation. Taken individually, perhaps the improvements are not very clear, but *collectively*, a trend does seem to appear in favor of syntactic coupling features generally, although we have not conducted formal statistical tests to validate this impression. A more detailed comparison between individual lines, inside the class of combinations that include MATRAX features, appears however difficult to make on the basis of the reported experiments.

References

Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. 2002. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(3):121–144.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2003. Syntax for Statistical Machine Translation: Final report of John Hopkins 2003 Summer Workshop. Technical report, John Hopkins University.

B. Roark, M. Saraclar, M. Collins, and M. Johnson. 2004. Discriminative language modeling with conditional random fields and the perceptron algorithm. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, July.

Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Éric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with non-contiguous phrases. In *HLT/EMNLP*.