# Translation reranking using source phrase dependency features

**Antonio Valerio Miceli-Barone**
Dipartimento di Informatica
Largo B. Pontecorvo, 3
56127 Pisa, Italy
miceli@di.unipi.it

## Abstract

We describe a N-best reranking model based on features that combine source-side dependency syntactical information and segmentation and alignment information. Specifically, we consider segmentation-aware ”phrase dependency” features.

## 1 Introduction

Dependency features have been used in the past for both direct translation and reranking (Gimpel and Smith, 2013), usually in a string-to-tree or a tree-to-tree configuration. These approaches generally require the decoder to be specifically designed to produce suitable dependency structures on its output, or to use a specialized target-side parser capable of parsing potentially ungrammatical and unidiomatic sentences.

Instead, we investigated a tree-to-string N-best reranking model suitable for use with a standard phrase-based decoder and a standard source-side dependency parser.

## 2 Source phrase dependency model

Dependency relations in a conventional dependency tree are syntactical relations between individual words. A phrase-based decoder, instead, operates in terms of phrase-pairs.

Each N-best candidate translation $e_i$ of a source sentence $f$ is defined by its derivation, which describes how $f$ has been segmented into source phrases, how these source phrases have been reoredered and for each source phrase which corresponding target phrase has been chosen.

In our model, we focus on the quality of phrase segmentation and reordering.

**Segmentation features** The source phrases produced by the segmentation performed by the decoder do not necessarily correspond to subtrees in the dependency parse tree (or forest) $g_f$ of the sentence. And if the dependency parse is not projective, subtrees do not necessarily correspond to contiguous phrases in any possible segmentation.

We propose a set of multiple features which operate at source phrase level, inspired by the concept of *phrase dependency* relations of Gimpel and Smith (2013):
Given a source phrase $\bar{f}_j$ in a derivation, we define the set of its parent phrases $PARENTS(\bar{f}_j)$ as the set of other phrases in the same derivation which contain at least one word that is a parent of some word in $\bar{f}_j$. We also define the sets of left parents $PARENTS_L(\bar{f}_j)$, right parents $PARENTS_R(\bar{f}_j)$, left children $CHILDREN_L(\bar{f}_j)$ and right children $CHILDREN_R(\bar{f}_j)$. Note that only word dependency relations that cross the phrase boundaries are relevant to the definition of these phrase dependency relations.

We propose the following segmentation phrase feature functions:
No parents $PARENTS(\bar{f}_j) = \emptyset$, no left par-

*Proceedings of SSST-9, Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 57–60,
Denver, Colorado, June 4, 2015. ©2015 Association for Computational Linguistics

ents $\mathit{PARENTS}_L(\bar{f}_j) = \emptyset$, no right parents $\mathit{PARENTS}_R(\bar{f}_j) = \emptyset$, one-sided parents $\mathit{PARENTS}_L(\bar{f}_j) = \emptyset \vee \mathit{PARENTS}_R(\bar{f}_j) = \emptyset$. Unambiguous (no more than one) parents $|\mathit{PARENTS}(\bar{f}_j)| \leq 1$, Unambiguous left parents $|\mathit{PARENTS}_L(\bar{f}_j)| \leq 1$, Unambiguous right parents $|\mathit{PARENTS}_L(\bar{f}_j)| \leq 1$. Unique parent $|\mathit{PARENTS}(\bar{f}_j)| = 1$. No children $\mathit{CHILDREN}(\bar{f}_j) = \emptyset$, no left children $\mathit{CHILDREN}_L(\bar{f}_j) = \emptyset$, no right children $\mathit{CHILDREN}_R(\bar{f}_j) = \emptyset$, one-sided children $\mathit{CHILDREN}_L(\bar{f}_j) = \emptyset \vee \mathit{CHILDREN}_R(\bar{f}_j) = \emptyset$.

When phrase segmentation breaks the syntactic structures these features should be able to detect it, and the model will penalize (or perhaps reward) different types of breakages using parameters automatically learned by tuning, similarly to Cherry (2008) or Marton and Resnik (2008).

**Distortion features** We consider pairs of source phrases which are aligned to target phrases that are contiguous in target order.

Let $\tilde{f}_j \equiv (\bar{f}_{a(j-1)}, \bar{f}_{a(j)})$ be one of such pairs. We define the following, mutually exclusive, feature functions:
Unique parent-child $\mathit{PARENTS}(\bar{f}_{a(j)}) = \{\bar{f}_{a(j-1)}\}$. Unique child-parent $\mathit{PARENTS}(\bar{f}_{a(j-1)}) = \{\bar{f}_{a(j)}\}$. Siblings with unique parent $\exists j' : \mathit{PARENTS}(\bar{f}_{a(j)}) = \mathit{PARENTS}(\bar{f}_{a(j-1)}) = \bar{f}_{j'}$. None of the above.

We also define the inversion feature function $a(j-1) > a(j)$ which is included both as an individual feature and in logical conjunction with each of the feature functions defined above, resulting in a total of nine boolean distortion feature functions.

These features detect reordering operations which swap syntactic structures related by a dependency relation between themselves or with a shared parent structure, similarly to the reordering operations in the *synchronous dependency insertion grammar* of Ding and Palmer (2005) or the *syntactic coupling* features of Nikoulina and Dymetman (2008).

**Scoring model** The feature functions defined in the two previous paragraphs are combined into a vector which is concatenated to the feature vector produced by the decoder and multiplied by a parameter vector $\theta$ to obtain the final reranking score for each candidate translation. $\theta$ is trained using a standard machine translation tuning technique, namely K-best batch MIRA (Cherry and Foster, 2012).

## 3 Experiments

**Setup** We tested our model in a Italian-to-English 1000-best translation reranking task.

We trained the baseline phrase-based system using a parallel corpus assembled from Europarl v7 (Koehn, 2005), JRC-ACQUIS v2.2 (Steinberger et al., 2006) and additional bilingual articles crawled from online newspaper websites[1], totaling 3,081,700 sentence pairs, which were split into a 3,075,777 sp. phrase-table training corpus, a 3,923 sp. tuning corpus, and a 2,000 sp. test corpus.

We trained and tuned phrase-based Moses (Koehn et al., 2007) using a ”sparse features” configuration (the ”word translation” and ”phrase translation” feature sets described by Chiang et al. (2009)). We performed model parameter tuning using k-best batch MIRA.
Non-projective dependency parse trees (actually, forests) for the Italian source sentences have been computed using the transition-based DeSR parser in tree revision configuration (Attardi and Ciaramita, 2007).

Significance was estimated using *paired bootstrap resampling* (Koehn, 2004).

**Results** The results of these experiments are shown in fig. 1.

We obtain a small but significant BLUE score improvement.

We also performed other experiments with slightly different feature function configurations but we obtained lower scores, although never lower than the baseline score of the decoder.

From a computational time point of view, the reranker adds a negligible overhead the the

[1] Corriere.it and Asianews.it

| Configuration | BLEU-c | BLEU |
|---|---:|---:|
| Moses + sparse feats. | 29.02 | 29.82 |
| Moses + sparse feats. + dep. feats. | 29.17 (+ 0.15) | 29.97 (+ 0.15) |

Figure 1: Experimental results. BLEU and case-insensitive BLEU scores over a 2,000 sp. it-en test corpus. Improvements are significant at the p ¡ 0.05 significance level.

runtime of the decoder, even in our unoptimized Python implementation.

**Conclusions and future work** We identified a set of syntactic dependency features which can provide small but significant translation quality improvements when used in N-best reranking, at least on the Italian-to-English language pair. We need to perform experiments on other language pairs to determine whether this result generalizes.

Spurious effects due to optimizer instability that can't be detected by our significance tests might be present. More advanced statistical tests such as Clark et al. (2011) should be performed to increase the confidence in the validity of our result.

In addition to reranking, our feature functions could also be used for decoding in a standard phrase-based or hierarchical translation system without a significant increase of decoding complexity, since they decompose additively over phrases or pair of phrase adjacent in target-order. Performing such experiment will be a natural extension of our work.

## References

Giuseppe Attardi and Massimiliano Ciaramita. 2007. Tree revision learning for dependency parsing. In Candace L. Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, *HLT-NAACL*, pages 388–395. The Association for Computational Linguistics.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436. Association for Computational Linguistics.

Colin Cherry. 2008. Cohesive phrase-based decoding for statistical machine translation. In *In Proceedings of ACL-08: HLT*, pages 72–80.

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

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics.

Yuan Ding and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 541–548, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin Gimpel and Noah A Smith. 2013. Phrase dependency machine translation with quasi-synchronous tree-to-tree features.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *ACL*, pages 1003–1011.

Vassilina Nikoulina and Marc Dymetman. 2008. Using syntactic coupling features for discriminating phrase-based translations (wmt-08 shared translation task). In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 159–162. Association for Computational Linguistics.

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