A Dependency Projection Model for Phrase-Based SMT

Christian Hardmeier

Uppsala University
Department of Linguistics and Philology
Box 635, 751 26 Uppsala, Sweden
christian.hardmeier@lingfil.uu.se

1. Introduction

Much of the information that is relevant for linguistic correctness in a sentence is best represented in hierarchical structures such as syntax trees. By contrast, the strongest models in statistical machine translation (SMT) represent sentences as linear sequences of words. This is especially true of phrase-based SMT (Koehn et al., 2003), an approach to SMT that delivers state-of-the-art performance for many language pairs, but treats sentences as mere sequences of words with no internal structure.

While phrase-based SMT is competitive with SMT approaches using context-free grammars to represent translation units (Chiang, 2007), it has notorious difficulties with linguistic features that depend on hierarchical relations such as number and gender agreement, especially if the two words that should agree with each other are separated by some intervening text. Here, we present some preliminary work on integrating syntactic knowledge into phrase-based SMT. Unlike previous work on syntactic language modelling such as that by Schwartz et al. (2011), we do not try to create parse trees for the machine translation output. Instead, we parse the input sentences with a dependency parser and use word alignments to link output words to input words. Our model then captures relations between the target language words corresponding to words engaged in particular source-side syntactic relations.

2. Dependency Projection Model

Our dependency projection model uses source-side dependency structure to model target-side relations between words. We start by parsing the input sentence with the MaltParser (Nivre et al., 2006). At decoding time, the model processes each source-language dependency arc in turn and identifies the words aligned to the head and modifier involved in the dependency relation by considering the word alignments stored in the SMT phrase table. In Figure 1, for instance, there is an *nsubjpass* arc connecting *dominated* to *production*. The head is aligned to the target word *dominée*, while the dependent is aligned to the set $\{production, de\}$. Based on this information, the model assigns a probabilistic score to each arc. The total model score of a document is equal to the sum of the logarithms of the individual arc scores. In dependency parsing parlance, our model is an arc-factored model.

The scoring model itself is a binary classifier trained to distinguish good examples from bad examples based on the features listed in Table 1. It is implemented as a feed-forward neural network whose architecture is depicted in

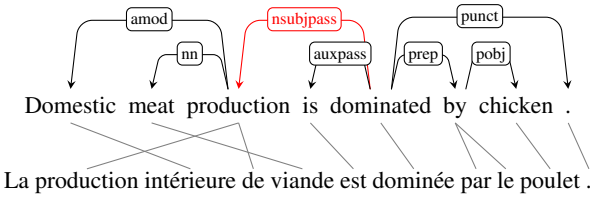

Figure 1: *Dependency projection model*

- head word
- head part-of-speech tag
- modifier word
- modifier part-of-speech tag
- target word set aligned to head
- target word set aligned to modifier
- dependency relation
- distance between head and modifier in the source language
- shortest distance between any pair of words in the aligned sets

Table 1: *Scoring model features*

Figure 2. The features pertaining to the source and the target representations of the head and the modifier, respectively, are fed into the network in the form of four feature vectors **Hs**, **Ht**, **Ms** and **Mt**. Each of these vectors is the concatenation of a part representing the words themselves and a part representing their part-of-speech (POS) tags. For the source elements, we use a one-hot representation with a single nonzero component for each part. The vectors of the target elements represent sets of words, so we use binary indicator features for each word and POS tag occurring in the set. These input vectors are projected onto lower-dimensional embedding vectors in the layer **E**. The embedding weights of the **Hs** and the **Ms** inputs are tied, and so are those of the **Ht** and **Mt** inputs. In addition to the word embeddings, the **E** layer has a number of components representing the distance and dependency relation features mentioned in Table 1. The **E** layer is then projected onto a hidden layer **H**, which is finally reduced to a single output value **O** corresponding to the score of an arc. All the layers have logistic sigmoid activation functions.

To train this network, we need a training set of good and bad examples. We generate positive training examples from an input text and a reference translation created by a human translator. We parse the source language text, align the source and the reference translation at the word level by concatenating them with a large chunk of parallel text and running the FastAlign algorithm (Dyer et al., 2013) and

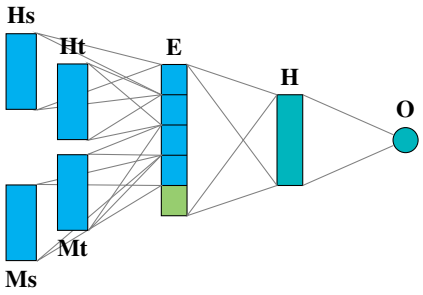

Figure 2: *Neural network architecture*

extract examples for every arc found. Negative examples are produced in a similar way, but using n-best lists output by a baseline machine translation system instead. To avoid generating negative examples from good choices made by the SMT system, we count examples as positive if they also occur in a reference translation of the same text. The neural network is trained to minimise cross-entropy with backpropagation and a variant of stochastic gradient descent.

We implement this model as a feature function in the Docent decoder (Hardmeier et al., 2012; Hardmeier et al., 2013). Docent implements a search procedure based on local search. At any stage of the search process, its search state consists of a complete document translation, making it easy for feature models to access the complete document with its current translation at any point in time. The search algorithm is a stochastic variant of standard hill climbing. At each step, it generates a successor of the current search state by randomly applying one of a set of state changing operations to a random location in the document, and accepts the new state if it has a better score than the previous state. Implemented operations include changing the translation of a phrase, changing the word order by swapping the positions of two phrase sequences, and resegmenting phrases. The initial state can either be generated randomly, or be based on an initial run from Moses. This setup is not limited by dynamic programming constraints, and enables us to use full sentence (and document) context to compute feature scores.

3. Experimental Observations

We ran some initial experiments with the English–French SMT system we submitted to the WMT 2014 shared task (Hardmeier et al., 2014), using only the standard sentence-level baseline features and the dependency projection model. The neural network has word embeddings of size 200 and an **H** layer of size 1000 and is regularised with an ℓ_2 weight cost of 10^{-5}. The source and target vocabularies are limited to the 2000 most frequent words. Less frequent items are mapped to a single item OTHER. The decoder is initialised with a Moses run including all features except for the dependency model. With a test set of news documents compiled from a number of WMT test sets, we do not observe any significant effect on BLEU scores. Manual inspection reveals that there are both positive and the negative effects on translation quality, and all effects are small, but some of the positive effects are quite interesting.

The most frequent improvement achieved by our model is the insertion of perfect tense auxiliary verbs omitted by the baseline model such as in the following example:

Kubatov admis avoir émis des factures...
Kubatov a admis avoir émis des factures...

In some examples, it improves word ordering in names, where the n-gram model fails because of unknown words:
L'historien de l'art Pierre Zurich-based Cornelius Claussen
L'historien de l'art Zurich-based Peter Cornelius Claussen
On the downside, the model tends to favour singular auxiliary verbs even with plural subjects. This can be explained by the fact that an arc-factored model does not have access to the necessary information to enforce subject-verb agreement and suggests that we should try using a second-order model over pairs of dependency arcs.

To conclude, our initial results suggest that dependency projection modelling shows some promise as a way to improve the syntactic structure of phrase-based SMT output. In future work, we plan to focus on improving the classifier to make more accurate predictions.

References

David Chiang. 2007. Hierarchical phrase-based translation. *Computational linguistics*, 33(2):201–228.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta (Georgia, USA).

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island (Korea).

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 193–198, Sofia (Bulgaria).

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, Aaron Smith, and Joakim Nivre. 2014. Anaphora models and reordering for phrase-based SMT. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 122–129, Baltimore (Maryland, USA).

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton (Canada).

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A language-independent system for data-driven dependency parsing. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC-2006)*, pages 2216–2219, Genoa (Italy).

Lane Schwartz, Chris Callison-Burch, William Schuler, and Stephen Wu. 2011. Incremental syntactic language models for phrase-based translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 620–631, Portland (Oregon, USA).