Learning to Prune: Context-Sensitive Pruning for Syntactic MT

Wenduan Xu, Yue Zhang, Philip Williams and Philipp Koehn
wenduan.xu@cl.cam.ac.uk, yue_zhang@sutd.edu.sg, p.j.williams-2@sms.ed.ac.uk, pkoehn@inf.ed.ac.uk

Abstract

We present a context-sensitive chart pruning method for CKY-style MT decoding. Source phrases that are unlikely to have aligned target constituents are identified using sequence labellers learned from the parallel corpus, and speed-up is obtained by pruning corresponding chart cells. The proposed method is easy to implement, orthogonal to cube pruning and additive to its pruning power. On a full-scale English-to-German experiment with a string-to-tree model, we obtain a speed-up of more than 60% over a strong baseline, with no loss in BLEU.

Problem

Syntactic MT models suffer from decoding efficiency bottlenecks. Especially for more expressive, linguistically-motivated syntactic MT models, where the grammar complexity has grown considerably over hierarchical phrase-based models and decoding still suffers from efficiency issues.

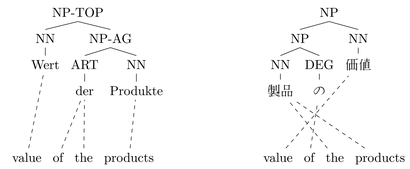
Method Overview

We study a chart pruning method for CKY-style MT decoding that is orthogonal to cube pruning and additive to its pruning power. The main intuition of our method is to find those source phrases (i.e. any sequence of consecutive words) that are unlikely to have any consistently aligned target counterparts according to the source context and grammar constraints.

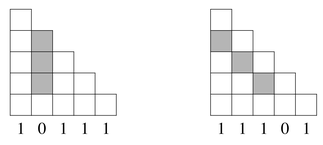
We call our method *context-sensitive pruning* (CSP); it can be viewed as a bilingual adaptation of similar methods in monolingual parsing (Roark and Hollingshead, 2008; Zhang et al.,2010) which improve parsing efficiency by "closing" chart cells using binary classifiers. Our contribution is that we demonstrate such methods can be applied to synchronous-grammar parsing by labelling the source-side alone. This is achieved through a novel training scheme where the labelling models are trained over the word-aligned bitext and gold-standard pruning labels are obtained by projecting target-side constituents to the source words. To our knowledge, this is the first work to apply this technique to MT decoding.

Pruning by Labelling

In contrast to monolingual parsing, our pruning decisions are based on the source context, its target translation and the mapping between the two. The key question is: how to inject target syntax and word alignment information into our labelling models; this is important since the syntactic correspondence between different language pairs is different. For example, "the products" does not have a consistent alignment on the target side on the left (en-de), while it does on the right (en-jp).



We use binary tags to indicate whether a source word can start or end a multi-word phrase that has a consistently aligned target constituent. Under this scheme, a b-tag value of 1 indicates that a source word can be the start of a source phrase that has a consistently aligned target phrase; similarly an e-tag of 0 indicates that a word cannot end a source phrase. If either the b-tag or the e-tag of an input phrase is 0, the corresponding chart cells will be pruned. The pruning effects of the two types of tags are illustrated below. In general, 0-valued b-tags prune a whole column of chart cells and 0-valued e-tags prune a whole diagonal of cells; and the chart cells on the first row and the top-most cell are always kept so that complete translations can always be found.



We build a separate labeller for each tag type using gold-standard b- and e-tags, respectively. The labellers are trained with maximum-entropy models (Curran and Clark, 2003; Ratnaparkhi, 1996), using features similar to those used for suppertagging for CCG parsing (Clark and Curran, 2004). During testing, in order to prevent overpruning, a tag value of 0 is assigned to a word only if its marginal probability is greater than a cut-off value θ.

Gold-standard Labelling

For each training sentence pair, gold-standard b-tags and e-tags are assigned separately to the source words. First, we initialize both tags of each source word to 0s. Then, we iterate through all target constituent spans, and for each span, we find its corresponding source phrase, as determined by the word alignment. If a constituent exists for the phrase pair, the b-tag of the *first* word and the e-tag of the *last* word in the source phrase are set to 1s, respectively.

Input forward alignment $A_{e\sim f}$, backward alignment $\hat{A}_{f\sim e}$ and 1-best parse tree τ for f
Output Tag sequences $\boldsymbol{b}$ and $\boldsymbol{e}$ for e

1: **procedure** TAG(e, f, τ, A, $\hat{A}$)
2: $\quad l \leftarrow |e|$
3: $\quad$ **for** $i \leftarrow 0$ to $l-1$ **do**
4: $\quad\quad \boldsymbol{b}[i] \leftarrow 0$, $\boldsymbol{e}[i] \leftarrow 0$
5: $\quad$ **for** $f[i', j']$ in τ **do**
6: $\quad\quad \boldsymbol{s} \leftarrow \{\hat{A}[k] \mid k \in [i', j']\}$
7: $\quad\quad$ **if** $|\boldsymbol{s}| \leq 1$ **then continue**
8: $\quad\quad i \leftarrow \min(\boldsymbol{s})$, $j \leftarrow \max(\boldsymbol{s})$
9: $\quad\quad$ **if** CONSISTENT(i, j, i', j') **then**
10: $\quad\quad\quad \boldsymbol{b}[i'] \leftarrow 1$, $\boldsymbol{e}[j'] \leftarrow 1$

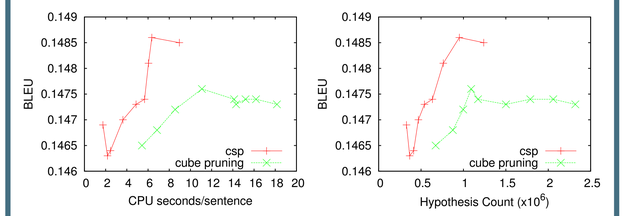
11: **procedure** CONSISTENT(i, j, i', j')
12: $\quad \boldsymbol{t} \leftarrow \{A[k] \mid k \in [i, j]\}$
13: $\quad$ **return** $\min(\boldsymbol{t}) \geq i'$ **and** $\max(\boldsymbol{t}) \leq j'$

Our definition of the gold-standard allows source-side labels to integrate bilingual information. On line 6, the target-side syntax is projected to the source; on line 9, consistency is checked against word alignment. From the gold standard data, we found 73.69% of the 54M words do not begin a multi-word aligned phrase and 77.71% do not end a multi-word aligned phrase; the 1-best accuracies of the two labellers tested on a held-out 20K sentences are 82.50% and 88.78% respectively.

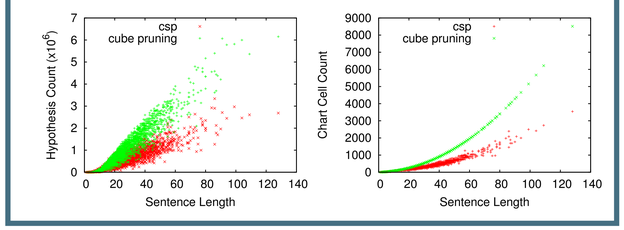
We use two independently trained labellers to perform b- and e-tag labelling separately prior to decoding. Training of the labelling models is able to complete in under 2.5 hours and the whole test set is labelled in under 2 seconds. A standard perceptron POS tagger trained on Wall Street Journal sections 2-21 of the Penn Treebank is used to assign POS tags for both our training and test data.

Results

A WMT Moses string-to-tree system is used as our baseline and decoding speed is measured by the average decoding time and average number of hypotheses generated per sentence. As shown below, the CSP decoder, which considers far fewer chart cells and generates significantly fewer subtranslations, consistently outperforms the slower baseline. It ultimately achieves a BLEU score of 14.86 at a probability cutoff value of 0.98, slightly higher than the highest score of the baseline.



The following two figures demonstrate the pruning power of CSP ($\theta = 0.95$) in comparison with the baseline (beam size = 300); across all the cutoff values and beam sizes, the CSP decoder considers 54.92% fewer translation hypotheses on average and the minimal reduction achieved is 46.56%.



References

[1] S. Clark, J.R. Curran: *The importance of su- pertagging for wide-coverage ccg parsing.* In Proc. COLING, 2004
[2] J.R. Curran, S. Clark: *Investigating gis and smoothing for maximum entropy taggers.* In Proc. EACL, 2003
[3] A. Ratnaparkhi: *A maximum entropy model for part-of-speech tagging.* In Proc. EMNLP, 1996
[4] Brian Roark, Kristy Hollingshead: *Classify- ing chart cells for quadratic complexity context-free inference*, In Proc. COLING, 2008
[5] Y. Zhang, B.G. Ahn, S. Clark, C. Van Wyk J.R. Curran, and L. Rimell: *Chart pruning for fast lexicalised-grammar parsing*, In Proc. COLING, 2010