Dependency Treelet Translation: Syntactically Informed Phrasal SMT

26 July 2018,
Jay Kejriwal
Index

• Introduction
• Dependency Treelet Translation
• Statistical models
• Decoding
• Optimization
• Results
Introduction

Translating domain-specific terminology and fixed phrases, grammatical generalizations are poorly captured and often mangled during translation.

Limitation of string based phrasal SMT

• reordering of words is limited
• generalization is poor for unseen phrases
Dependency Treelet Translation

The approach was developed by Chris Quirk, Arul Menezes and Colin Cherry in 2003 at Microsoft Research.

The propose the novel dependency treebased approach to phrasal SMT which uses treebased ‘phrases’ and a tree-based ordering model in combination with conventional SMT models to produce state-of-the-art translations.
Dependency Treelet Translation

The approach uses

• A source-language dependency parser
• A target language word segmentation component
• An unsupervised word alignment component to learn treelet translations
• Parallel sentence-aligned corpus
Dependency Treelet Translation

Parsing and alignment

• source language dependency parser that produces unlabeled, ordered dependency trees and annotates each source word with a part-of-speech (POS)

```
  startup properties and options
  Noun    Noun           Conj    Noun
```

Figure 1: An example dependency tree

Note: GIZA++ is used to obtain word alignments
Dependency Treelet Translation

Projecting dependency trees

- The dependency parsed sentences of the source are then projected onto the word-aligned parallel texts to produce word-aligned parallel dependency corpus.

Mappings

- One-to-one mappings: the projections are simple and the target tree becomes isomorphic to the source.
- In many-to-one alignments, multiple source words that are linked in the tree get projected onto a single target word.
- One-to-many alignments, a single source word corresponds to several target words that are contiguous in the tree.
Dependency Treelet Translation

Projecting dependency trees

• The system projects the source node to the rightmost word of the target phrase.

• The words that are unaligned in the target sentence are attached to the closest lower node to the left [or right] in the dependency tree.

(startup properties and options)

(propriétés et options de démarrage)

(a) Word alignment.
Dependency Treelet Translation

Projecting dependency trees

• If all the nodes to the left [or right] of a word $w_j$ are unaligned, the word is attached to the leftmost [or right-most] word that is aligned

(b) Dependencies after initial projection.
Dependency Treelet Translation

Projecting dependency trees

• The resulting target dependency tree may not be in the same surface ordering so reattachment pass, which reattaches each wrong node to the lowest of its ancestors
Dependency Treelet Translation

Extracting Treelet pairs

• Individual treelets are then extracted from this corpus, which together form a treelet translation model of source and target translation pairs.

• treelet is defined as an arbitrary connected subgraph of the aligned parallel dependency tree.

• The frequency count for the treelets is also maintained, which is then used for maximum likelihood estimation.

• A threshold is used to limit the size of the treelet, so as to limit the possible combination.
Statistical models

There are a number of models possible to build a statistical model for the phrasal SMT system. Here we discuss a few:

- Order Model
- Channel Model
- Target Model
Statistical models

Order Model

• treelet system incorporates syntactic information for ordering the phrases.

• It assigns a probability to the order of a target treelets given the sequence of source treelets.

• It is simplified by an assumption that phrases move as a whole, which predicts the probability of each given ordering of modifiers independently.
Statistical models

Order Model

- This can be represented as

\[ P(\text{order}(T) \mid S,T) = \prod_{t \in T} P(\text{order}(c(t)) \mid S,T) \]

where, S and T are the source and target treelet sequences respectively, and c is the function returning the list of nodes modifying t.
Statistical models

• Further, it is assumed that the position of each child can be modelled independently in terms of a head-relative position:

• Small set of features reflecting very local information in the dependency tree to model $P(\text{pos}(m,t) \mid S, T)$:

• Lexical items head, modifiers, part of speech are considered

$$P(\text{order}(T) \mid S, T) = \prod_{m \in c(t)} P(\text{pos}(m,t) \mid S,T)$$
Statistical models

Example

\[ P(\text{pos}(m1) = -1 \mid \text{lex}(m1) = "la", \text{lex}(h) = "propriété", \text{lex}(\text{src}(m1)) = "the", \text{lex}(\text{src}(h)) = "property", \text{cat}(\text{src}(m1)) = \text{Determiner}, \text{cat}(\text{src}(h)) = \text{Noun}, \text{position}(\text{src}(m1)) = -2) \times \]

\[ P(\text{pos}(m2) = +1 \mid \text{lex}(m2) = "Cancel", \text{lex}(h) = "propriété", \text{lex}(\text{src}(m2)) = "Cancel", \text{lex}(\text{src}(h)) = "property", \text{cat}(\text{src}(m2)) = \text{Noun}, \text{cat}(\text{src}(h)) = \text{Noun}, \text{position}(\text{src}(m2)) = -1) \]
Statistical models

Channel Model

• The system uses two distinct channel models, a maximum likelihood estimate (MLE) model and one using IBM Model-1 word-to-word alignment probabilities.

• The MLE is effective in modelling idioms and other non-literal translation of phrases but suffers from data sparseness since these occur very rarely in the corpus.

• Word-to-word model is biased towards literal translations and is not as effective for idiomatic uses.
Statistical models

**Target Model**

• Given an ordered target language dependency tree, it is trivial to read off the surface string

• For improving the fluency of the translation, it uses a trigram language model with Kneser-Ney smoothing.
Decoding

Unlike string-based approaches, which use simple left-to-right decoding, the decoding of treelet based approach is more complicated as it uses bottom up, exhaustive search

Steps for decoding:
• Employ treelet translation pairs instead of single word translations.
• Instead of modelling rearrangements as either preserving source order or swapping source order, we allow the dependents of a node to be ordered in any arbitrary manner
• Use the order model described to estimate probabilities.
• Finally, use a log-linear framework for model combination that allows any amount of other information to be modelled.
Decoding

Bottom up approach

Moving bottom up through the input tree, a list of candidate translations are computed for the input sub tree rooted at each node. Each such interleaving is scored using the models previously described and added to the candidate translation list for that input node. The resultant translation is the best scoring candidate for the root input node.
Figure 3: Example decoder structures.
Optimization

• **N-best lists:** Instead of keeping the full list of translation candidates for a given input node, keep a topscoring subset of the candidates.

• **Pruning treelet translation pairs:** The Channel model scores and treelet size are powerful predictors of translation quality defining ratio and threshold.

• **Greedy ordering:** Limiting the number of ordering possibilities it considers each potential modifier position in turn, greedily picking the most probable child for that slot, moving on to the next slot, picking the most probable among the remaining children for that slot and so on.
Evaluation (Training)

1. A parallel English-French corpus containing 1.5 million sentences of Microsoft technical data

2. Training data ranging from 1,000 to 300,000 sentences and 250 sentences for lambda training.

3. 2,000 sentences for development testing & parameter tuning and 10,000 sentences for testing
Evaluation(Results)

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU Score</th>
<th>Sents/min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharaoh monotone</td>
<td>37.06</td>
<td>4286</td>
</tr>
<tr>
<td>Pharaoh</td>
<td>38.83</td>
<td>162</td>
</tr>
<tr>
<td>MSR-MT</td>
<td>35.26</td>
<td>453</td>
</tr>
<tr>
<td>Treelet</td>
<td>40.66</td>
<td>10.1</td>
</tr>
</tbody>
</table>

Figure 4: System Comparison.

- Note: Pharaoh monotone : phrase reordering disabled
Reference


Thank You