Rule-based preprocessing in a statistical system

Anna Bródy
Hybrid Machine Translation
2018
Preprocessing

- Tokenization, lemmatisation, NER
- Spelling errors, incomplete sentence pairs
- Statistical methods in RBMT systems
- Rules for SMT systems
  - Automatically extracted rules (with non-linguistics or with syntactic knowledge)
  - Manually written rules
What is preprocessing?

- Tokenization, lemmatisation, NER
- Spelling errors, incomplete sentence pairs
- Statistical methods in RBMT systems
- Rules for SMT systems
  - Automatically extracted rules (with non-linguistics or with syntactic knowledge)
  - Manually written rules → Hybrid MT
Clump-based baseline SMT with rewrite patterns

Why is SMT not enough?
  - Word-based < Phrase-based
  - BUT:
    - Bad phrase order
    - Reorderings do not keep phrase boundaries
Xia & McCord, 2004

- Clump-based baseline SMT with rewrite patterns
- Why is SMT not enough?
  - Word-based < Phrase-based
  - BUT:
    - Bad phrase order
    - Reorderings do not keep phrase boundaries
    - → Use rewrite rules to reorder the words in source sentences to mimic target language, use these as input in an SMT system
Xia & McCord, 2004 - Baseline SMT system

- **Training**
  - words alignment in every sentence pair
  - clump pairs are extracted
  - clump pairs and their countr form the clump dictionary
(a) word alignment

France is the first western country

la france est le premier pays occidental

(b) some of the clump pairs

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>france, la france</td>
</tr>
<tr>
<td>is</td>
<td>est, first, the first</td>
</tr>
<tr>
<td>is the first</td>
<td>est le premier</td>
</tr>
<tr>
<td>first western country</td>
<td>premier pays occidental</td>
</tr>
<tr>
<td>western country</td>
<td>pays occidental</td>
</tr>
</tbody>
</table>
Xia & McCord, 2004 - Baseline SMT system

- Decoding
  - segments source sentences into clumps
  - translates each one with dictionary
  - all translations are ranked
  - the one with the highest score "wins"
  - non-monotonic translation order to account for the word order differences
Xia & McCord, 2004 - Baseline SMT system

- Problems - two limitations
  - sentences with identical syntactic structure are not recognized
    – additional work for the system that could be avoided
  - clumps that are not phrases could be reordered in a way that is
    results in ungrammaticalilty – system again doesn’t recognize
    this
Xia & McCord, 2004 - Hybrid approach

- General idea
  - learn rewrite patterns from training data
  - use them to reorder source sentences before the decoding phase
  - monotonic translation of reordered clumps
Xia & McCord, 2004 - Hybrid approach

- **Training**
  - learn rewrite patterns
    - parsing
    - phrase alignment
    - pattern extraction
  - reorder source sentence with patterns
  - train baseline system with reordered source and target data to create the clump dictionary
Xia & McCord, 2004 - Hybrid approach

- Decoding
  - reorder source sentences
  - monotonic translation
Xia & McCord, 2004 - Rewrite patterns

- patterns are quintuples
  - (SrcRule, TrtRule, SrcHeadPos, TgtHeadPos, ChildAlign)
Xia & McCord, 2004 - Rewrite patterns

- patterns are quintuples
  - \((\text{SrcRule}, \text{TrtRule}, \text{SrcHeadPos}, \text{TgtHeadPos}, \text{ChildAlign})\)
  - \(\text{SrcRule}/\text{TgtRule}: I(X) \rightarrow I(X_1) \ldots I(X_m)\)
    - \(I(X)\): label of a node in the tree (any kind of syntactic label)
    - \(I(X_1) \ldots I(X_m)\): child of \(X\)
  - \(\text{SrcHeadPos}/\text{TgtHeadPos}\): an integer giving the position of \(X\)'s head in \(X_1 \ldots X_m\)
patterns are quintuples

- \( (\text{SrcRule}, \text{TrtRule}, \text{SrcHeadPos}, \text{TgtHeadPos}, \text{ChildAlign}) \)
- \( \text{SrcRule/TgtRule}: l(X) \rightarrow l(X_1) \ldots l(X_m) \)
  - \( l(X) \): label of a node in the tree (any kind of syntactic label)
  - \( l(X_1) \ldots l(X_m) \): child of \( X \)
- \( \text{SrcHeadPos/TgtHeadPos}: \) an integer giving the position of \( X \)'s head in \( X_1 \ldots X_m \)
- \( \text{ChildAlign}: \) injective (but not bijective) correspondence between \( X_i \) and \( Y_i \)
Xia & McCord, 2004 - Rewrite patterns

- patterns are quintuples
  - \((\text{SrcRule}, \text{TrtRule}, \text{SrcHeadPos}, \text{TgtHeadPos}, \text{ChildAlign})\)
  - \text{SrcRule/TgtRule}: \(l(X) \rightarrow l(X_1) \ldots l(X_m)\)
    - \(l(X)\): label of a node in the tree (any kind of syntactic label)
    - \(l(X_1) \ldots l(X_m)\): child of \(X\)
  - \text{SrcHeadPos/TgtHeadPos}: an integer giving the position of \(X\)'s head in \(X_1 \ldots X_m\)
  - \text{ChildAlign}: injective (but not bijective) correspondence between \(X_i\) and \(Y_i\)

- a rewrite pattern is lexicalized if at least one source node label \(X_i\) includes the head word \(\rightarrow\) when head word can influence word order
1. Parsing source and target (optional) sentences of bilingual text
Slot Grammar parsers (dependency oriented system)
1. Parsing source and target (optional) sentences of bilingual text
2. Aligning phrases
Word alignment using a word aligner trained on bilingual data

- For each phrase $S$ and $T$, calculate the percentage of words in $S$ and $T$ that are linked together

$$\text{Score}(S,T) = \frac{\# \text{Links}(S,T)}{\text{Span}(S) + \text{Span}(T)}$$

- Each word is identified by the node number of the phrase headed by the word, which is the word’s position number in the sentence
1. Parsing source and target (optional) sentences of bilingual text
2. Aligning phrases
3. Extracting rules
given a pair of SrcTree and TgtTree and a phrase alignment Al (the set of & target node pairs), extract all rewrite patterns of the form $X_1 \ldots X_m \rightarrow Y_1 \ldots Y_m$ iff

- $X_i$s are siblings in SrcTree, $Y_i$s are siblings in TgtTree
- parent node $X$ of $X_i$ aligns with parent node $Y$ of $Y_i$ – injective correspondence of $X_i$ and $Y_i$
- $X_i$ must contain the head of $X$ and $Y_i$ must contain the head of $Y$ and the heads must align in Al
- and aligned pair of $X_i$ and $Y_i$ in Al is either both lexicalized or unlexicalized in the pattern

- only generate rules whose source rule length is not longer than 5
Unlexicalized rules:
1: $NP_0 \lor NP_1 \rightarrow NP_0 \lor NP_1$
2: $N \rightarrow Det \; N$
3: $Adj_1 \; Adj_2 \; N \rightarrow Adj_1 \; N \; Adj_2$
4: $Adj \; N \rightarrow Adj \; N$
5: $Adj \; N \rightarrow N \; Adj$

Lexicalized rules:
6: $Adj(first) \; N \rightarrow Adj(premier) \; N$
7: $Adj \; N(country) \rightarrow Adj \; N(pays)$
8: $Adj(western) \; N \rightarrow N \; Adj(occidental)$
9: $Adj \; N(country) \rightarrow N(pays) \; Adj$
1. Parsing source and target (optional) sentences of bilingual text
2. Aligning phrases
3. Extracting rules
4. Organizing the patterns
Xia & McCord, 2004 - Rewrite patterns - Learning

- millions of patterns, many of them lexicalized
- merge identical patterns
- two steps:
  - put rules in the same group if their source rule is identical & normalize counts & default pattern for the source rule is with the highest count
  - for each group pair \((G_i, G_j)\), add a link from \(G_i\) to \(G_j\) iff there is no other group inbetween and the source rule of the latter is more \textit{specific} than the first, i.e. a lexicalized version of the same rule or a superstring
a pattern group is applicable to a node, if the context-free rule representation of the node is identical to or more specific than the source rule of the pattern group.

given a tree and a rewrite pattern hierarchy, we apply the pattern on the tree by:

- going through each internal node
- applying the most specific applicable pattern to reorder the children of the node
- we get the new tree with the target language order.
Xia & McCord, 2004 - Experiments

- 90-million-word English-French Canadian Hansard corpus for training of the baseline and the hybrid system
- 56 thousand patterns
- TestSet1 - 3971 sentences from Hansard (not included in training)
- TestSet2 - 500 sentences from news articles
- 10% relative improvement in Bleu measure