The ILLC-UvA SMT System for IWSLT 2010

Maxim Khalilov and Khalil Sima'an

{m.khalilov, k.simaan}@uva.nl
Institute for Logic, Language and Computation
Universiteit van Amsterdam
Amsterdam, The Netherlands

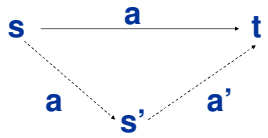
Abstract

In this study we give an overview of the ILLC-UvA (Institute for Logic, Language and Computation - University of Amsterdam) submission to the IWSLT 2010 evaluation campaign. It outlines the architecture and configuration of the novel features we are introducing: a syntax-based model for source-side reordering via tree transduction and accurate training data selection.

We have concentrated on the Chinese-to-English and English-to-Chinese DIALOG translation tasks.

Reordering as a preprocessing step

Source reordering of *s* is as successful as much as the alignment *a'* between the resulting permutation *s'* and *t* is monotone. Our task is a to model learning from a word-aligned parallel corpus (*s*--> *s*') a model of source permutation from *s* to *s'*, where *s'* has as monotone alignment with *t* as possible.



Syntactic transfer as a source permutation

We define source permutation as the problem of learning how to *transfer* a given source parse-tree into a parse tree that minimizes the divergence from target word-order.

We model the tree transfer $\tau_s \rightarrow \tau_{s'}$ as a sequence of local, independent transduction operations, each transforming the current intermediate tree τ_{s_i} into the next intermediate tree $\tau_{s_{i+1}}$, with $\tau_{s_0} \rightarrow \tau_s$ and $\tau_{s'_n} \rightarrow \tau_{s'}$.

A reordering model via tree transduction

1. Our model aims at learning from the source permuted parallel corpus $s \rightarrow s'$ a probabilistic optimization $\arg\max_{\pi(s)} P(\pi(s) \mid s, \tau_s)$

2. We view the set of permutations $s \rightarrow s'$ as a sequence of local tree transductions $\tau_{s'_0} \rightarrow \ldots \rightarrow \tau_{s'_n}$, where $s'_0 \rightarrow s$ and $s'_n \rightarrow s'$, and each transduction $\tau_{s'_{i-1}} \rightarrow \ldots \rightarrow \tau_{s'_i}$ is defined using a tree transduction operation.

3. Local transduction $\tau_{s'_{i-1}} \rightarrow \ldots \rightarrow \tau_{s'_i}$ is modeled by a single-node operation permuting the ordered sequence of children α_x dominated by node x.

4. We use an intersection of *s-to-t* and *t-to-s* word alignments, leaving the unique *1-to-1* links that correspond to the strongest lexical weights.

We take a pragmatic approach greedily selecting the single most likely local transduction at every intermediate point $\tau_{s'_{i-1}} \rightarrow \ldots \rightarrow \tau_{s'_i}$, using the following formula:

$$P\left(\pi(\alpha_x) \mid p.node\ x \rightarrow \alpha_x, tree\ context\ of\ x\right)^{\alpha} \times \left(\frac{P(s'_{i-1})}{P(s'_i)}\right)^{\beta}$$

- The conditional probability $P\left(\pi(\alpha_x) \mid p.node\ x \rightarrow \alpha_x, tree\ context\ of\ x\right)$ is estimated using a Maximum Entropy framework where features are defined to capture the permutation as a class.
- The string probability $P(s'_i)$ is estimated using an *n*-gram language model trained on the s' side of the source permuted corpus $s \rightarrow s'$

Tree context (features)

- Local tree topology: node + its children
- Dependency features: (1) the POS tag of the head word of the current node, (2) the sequence of POS tags of the head words of its child nodes
- Syntactic features (binary): (1) whether the parent node is a child of the node annotated with the same syntactic category, (2) whether the parent node is a descendant of the node annotated with the same syntactic category

Training data selection

1. The concatenation of BTEC and DIALOG corpora is used for training.

2. Individual phrase tables are extracted for sets with coinciding number of references. We then used the Moses capability to use of multiple translation tables during decoding:
   - Translation options are collected from one table, and additional options are collected from the other tables,
   - There is an additional table that consists of the intersection of the initial phrase tables, shared phrase pairs are removed from initial tables.

3. The target-side language model trained on the concatenation of the DIALOG and BTEC corpora to select a single best reference. Then, the selected references and its source counterpart are concatenated with the training corpus.

System	BLEY dev	BLEU test
Train only	42.16	33.39
Phrase tables	43.06	35.15
Phrase-tables+intersection	42.73	32.35
Train+best references	42.73	36.07

Experiments and submissions

System	α	β	LM order	Data selection	Submission	BLEU dev	BLEU test
Chinese-to-English							
1	0	0	-	Train+ best reference	Primary	42.73	36.07
2	0	1	-		-	41.80	35.03
3	1	0	3		-	41.84	35.17
4	1	0	4		Secondary1	42.12	35.61
5	1	1	4		Secondary2	41.91	35.61
English-to-Chinese							
1	0	0	-	Train+ best reference	Secondary1	29.15	32.17
2	0	1	-		-	28.52	31.84
3	0	1	-		-	28.15	31.93
4	1	1	3		Primary	29.91	32.76
5	1	1	4		Secondary2	28.65	32.15

- ✓ Baseline: Moses-based + 3-gram target-side LM
- ✓ MSD reordering model
- ✓ Stanford parser
- ✓ Le Zhang's maxent toolkit