Introduction

Under-resourced language pair: Scarcity of parallel corpora

SMT Problem:
No direct data → no SMT training
Insufficient data → poor SMT performance

Pivoting involves the use of *another language* to include resources available.

E.g. English to Slovak via Czech, Vietnamese to Czech via English

Pivoting Methods

System Cascades one system after another

Synthetic Corpus translates the pivot side of a corpus

Phrase Table Triangulation combines two phrase tables: source-pivot and pivot-target

Motivation

Promising results reported using phrase table triangulation, but no open-source tool

We decided to fill the gap and implement an easy-to-use tool.

Pivoting - It’s an MT thing

It is NOT the *pivot* method, which aims to balance the IR scores by the document length
It is NOT the *pivot* approach to cross lingual information retrieval, closer but still NO.

Phrase Table Triangulation Method

Linking Source and Target Phrases by connecting $s$ and $t$ whenever there exists a pivot phrase $p$ such that $s\,p$ is listed in the source-pivot and $p\,t$ is listed in the pivot-target phrase table.

Word Alignment for Linked Phrases by tracing the alignments from each source word $s \in s$ over any pivot word $p \in p$ to each target word $t \in t$.

Feature Values for Constructed Phrase Pairs:

Pivoting Probabilities

Both phrase and lexical probs merged:

a) assuming independence [sum]

b) by using the most prominent sense [max]

Pivoting Co-Occurrence Counts

1) Take min/max/mean ($f$) of each count

2) Estimate probabilities as usual:

Experiments

Our Experiment:

Results of triangulation are comparable but not better than the direct system

Improvement made by merging direct and pivoted phrase tables (Moses toolkit available)

Importance: different languages, domains and corpora may show different behavior patterns.

<table>
<thead>
<tr>
<th>Method</th>
<th>Table Size [#/pairs]</th>
<th>vi→cs BLEU</th>
<th>cs→vi BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct System</td>
<td>8.8M</td>
<td>7.62</td>
<td>10.59</td>
</tr>
<tr>
<td>Best Pivoted System</td>
<td>61.5M</td>
<td>7.44</td>
<td>10.28</td>
</tr>
<tr>
<td>Combination 1 (Linear Interpolation)</td>
<td>69.3M</td>
<td>8.33</td>
<td>11.98</td>
</tr>
<tr>
<td>Combination 3 (Alter. Decoding Paths)</td>
<td>8.8M/61.5M</td>
<td>8.34</td>
<td>11.85</td>
</tr>
</tbody>
</table>

Conclusion

Contact

TmTriangulate is freely available here:
https://github.com/tamhd/MultiMT

If you have any comments/suggestions, please send us an email to tamhd990 AT gmail DOT com

This work was supported by the grants no 644542 (Q21) and no 644402 (HimL) of the EU and SVV 260 104 of the Czech Republic. We used language resources hosted by the LINDAT/CLARIN project LM2010013 of the Ministry of Education, Youth and Sports.