TmTriangulate: A Tool for Phrase Table Triangulation

Tam Hoang, Ondřej Bojar
https://github.com/tamhd/MultiMT

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.

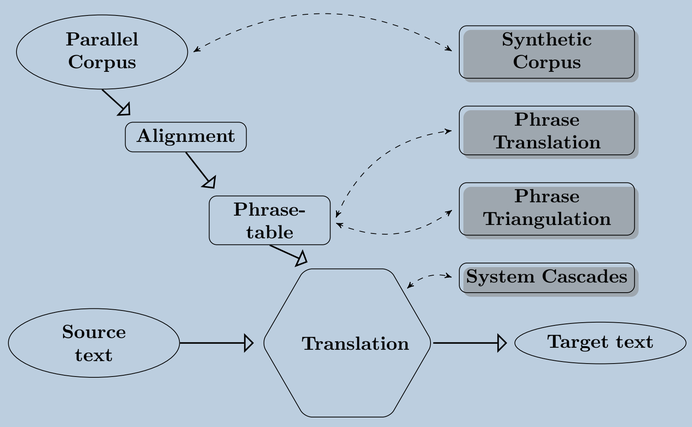
E.g: English to Slovak via Czech, Vietnamese to Czech via English

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.

Contact

`TmTriangulate` is freely available here:

`https://github.com/tamhd/MultiMT`

If you have any comments/suggestions, please send us an email to `tamhd1990 AT gmail DOT com`

Conclusion

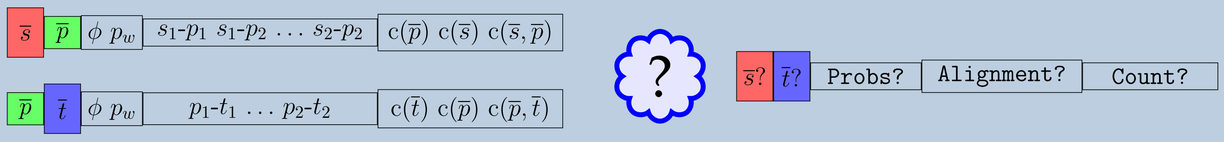
Our Experiment:

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

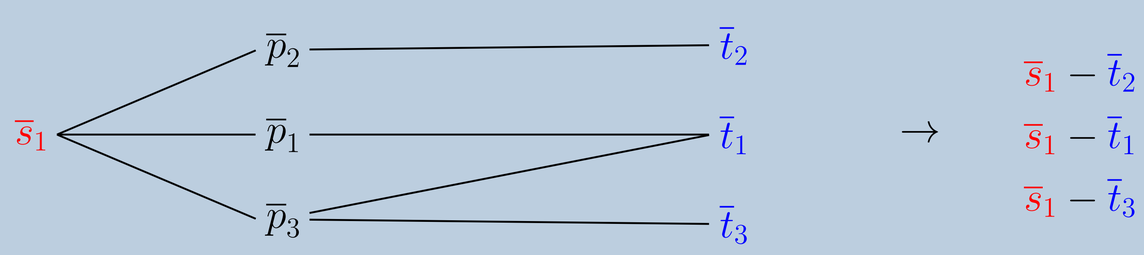
Improvement made by merging direct and pivoted phrase tables (Moses toolkit available)

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

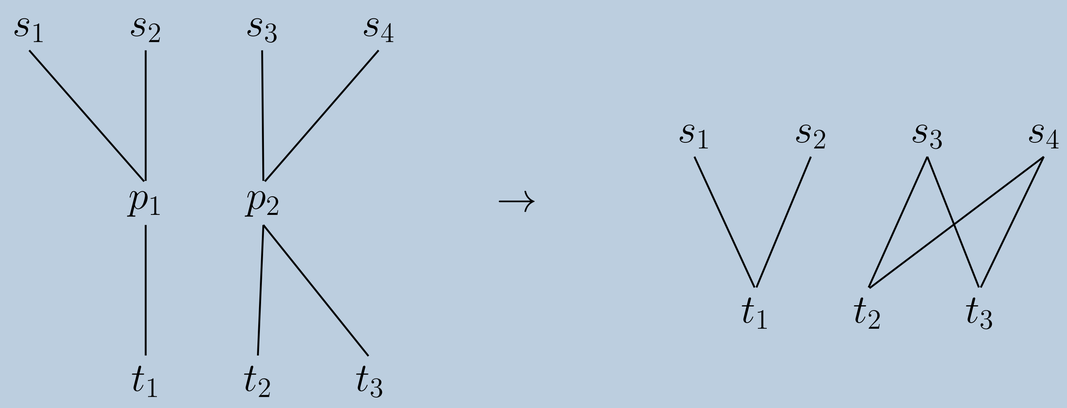
Phrase Table Triangulation Method



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



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



Feature Values for Constructed Phrase Pairs:

`Pivoting Probabilities` — `Pivoting Co-Occurrence Counts`

Both phrase and lexical probs merged:
a) assuming independence [sum]
b) using the most prominent sense [max]

$$\phi(\bar{s}|\bar{t}) \approx \sum_{\bar{p}} \phi(\bar{s}|\bar{p})\ \phi(\bar{p}|\bar{t})$$
$$\approx \max_{\bar{p}} \phi(\bar{s}|\bar{p})\ \phi(\bar{p}|\bar{t})$$
$$p_w(\bar{s}|\bar{t}) \approx \sum_{\bar{p}} p_w(\bar{s}|\bar{p})\ p_w(\bar{p}|\bar{t})$$
$$\approx \max_{\bar{p}} p_w(\bar{s}|\bar{p})\ p_w(\bar{p}|\bar{t})$$

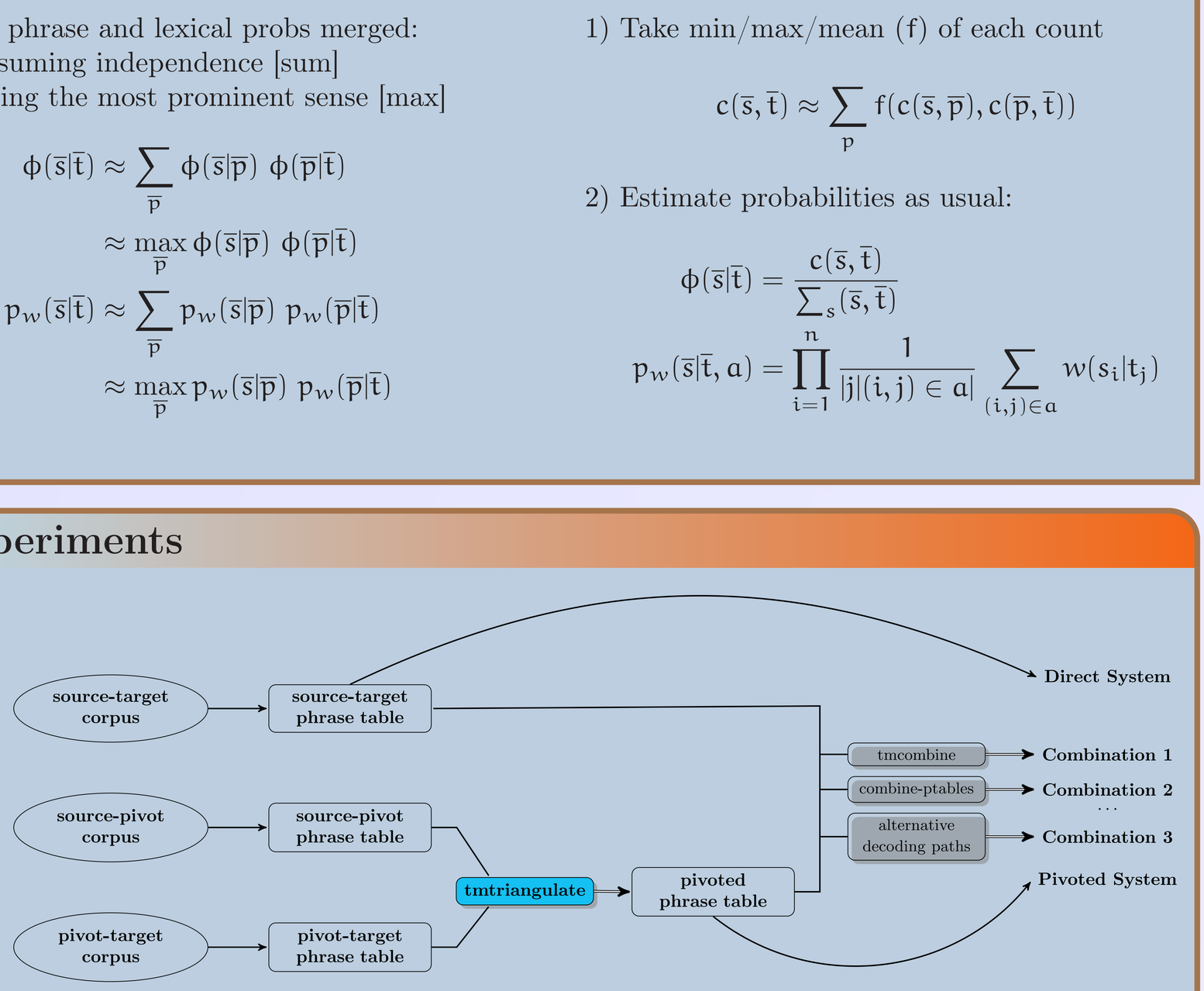
1) Take min/max/mean (f) of each count

$$c(\bar{s},\bar{t}) \approx \sum_{p} f(c(\bar{s},\bar{p}), c(\bar{p},\bar{t}))$$

2) Estimate probabilities as usual:

$$\phi(\bar{s}|\bar{t}) = \frac{c(\bar{s},\bar{t})}{\sum_s (\bar{s},\bar{t})}$$
$$p_w(\bar{s}|\bar{t},a) = \prod_{i=1}^{n} \frac{1}{|j|(i,j) \in a|} \sum_{(i,j)\in a} w(s_i|t_j)$$

Experiments



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

This work was supported by the grants n° 645452 (QT21) and n° 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.