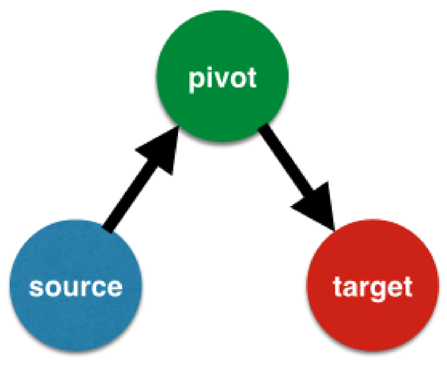
Supervised Phrase Table Triangulation for Low-Resource Languages

Tomer Levinboim, David Chiang
University of Notre Dame
{levinboim.1, dchiang}@nd.edu

Low-Resource MT

- **Goal**: Construct an MT system for a new language pair.
- **Problem**: Most language-pairs either have no parallel data or a very limited corpus.

Phrase Table Triangulation



- **Triangulation** approximates source-target translations from source-pivot and pivot-target tables:

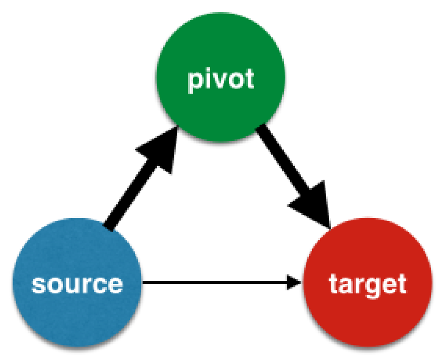
$$\hat{w}(t \mid s) = \sum_{p} w(t \mid p) \cdot w(p \mid s) \qquad (1)$$

- The **lexical weight** of an aligned phrase pair $(\mathbf{s}, \mathbf{t})$ is:

$$\hat{lex}(\mathbf{t} \mid \mathbf{s}) \propto \prod_{s \in \mathbf{s}} \sum_{t:(s,t) \in \mathbf{a}} \hat{w}(t \mid s)$$

where $\mathbf{a}$ denotes the word alignment.

Phrase Table Interpolation



- Given a source-target table w, we can improve upon triangulation using **interpolation**:

$$\alpha w + (1 - \alpha)\hat{w}$$

Analysis

1. Triangulation yields **noisy tables**
2. $\hat{w}$ and w are estimated **independently**
3. Interpolation can only improve words in the **intersection** of $\hat{w}$ and w

Research Question:
Can we improve on the triangulated / interpolated lexical weights using very limited source-target bilingual data?

Monolingual data

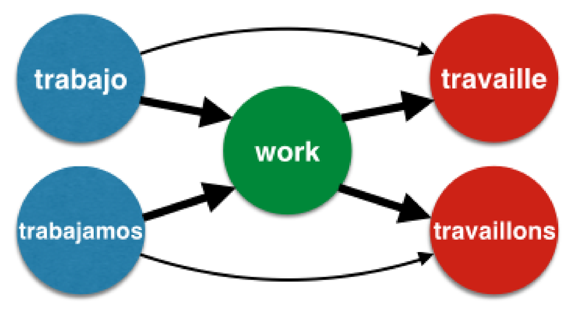
	Spanish	French	Malagasy
Tokens	1.5G	1.5G	58M

Monolingual data; Used to produce word embedding

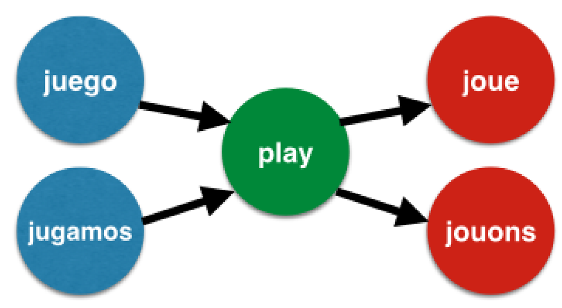
Experiment

1. Produce counts c^{sup} from bilingual data
2. Split word-pairs to 70/30% train/dev sets and train the model, selecting the Total Variation minimizing iteration on the dev set
3. Re-score and append lexical weights to the triangulated / interpolated phrase table

Conflated Translations in Triangulation



Interpolation may correct conflated translations.



Can we learn to correct conflations for word pairs unseen in source-target data?

- The independence assumption leads to many conflated translations
- Further conflations occur when pivoting through English, due to its poor morphology

Our approach:

1. Derive word co-occurrence counts $c^{\mathrm{sup}}(s, t)$ from limited source-target parallel data / dictionary
2. Train a discriminative model: correct translation (according to c^{sup}) should become likely, while incorrect ones should be down-weighted
3. Generalize beyond the intersection of w and $\hat{w}$, by representing words with embedding (word2vec)

Our Model

Let $q(t \mid s)$ denote the probabilities to be learned:

$$q(t \mid s) = \frac{1}{Z_s} \exp\left(v_s^T A v_t + \hat{w}_{st} \cdot h\right)$$

$$Z_s = \sum_{t \in \mathcal{T}(s)} \exp\left(v_s^T A v_t + \hat{w}_{st} \cdot h\right)$$

- v_s and v_t are source/target embeddings
- $\hat{w}$ is the triangulated word translation score

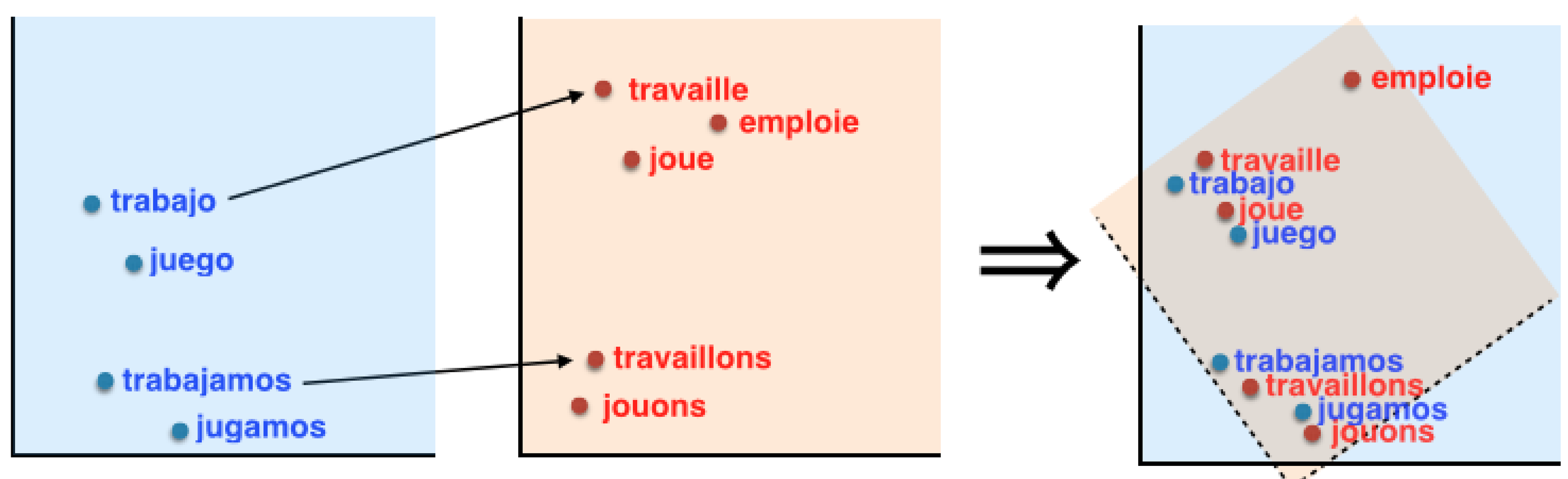
Optimization Problem

Learn the matrix A and scalar h:

$$\max_{A,h} L(A, h) = \max_{A,h} \sum_{s,t} c^{\mathrm{sup}}_{st} \log q(t \mid s).$$

- The matrix A maps between the source/target embedding spaces; h quantifies how much the triangulation $\hat{w}$ should be trusted
- Solved with gradient descent and AdaGrad

Learning a Discriminative Mapping



The matrix A should map correct Spanish-French translations closer together, and incorrect ones, such as trabajo-emploie, further apart

Learning from a 50k Spanish-French corpus

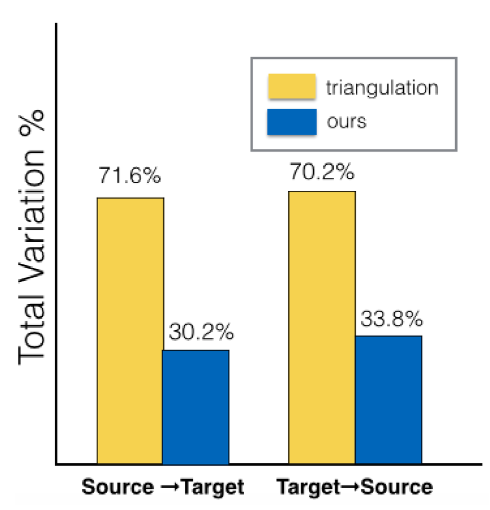
aceptamos	acceptons	accepter	croyons
norm. c^{sup}	1.0	-	-
$\hat{w}$ (triang.)	0.05	0.39	-
ours (no $\hat{w}$)	0.62	-	0.06
ours	0.98	0.01	-

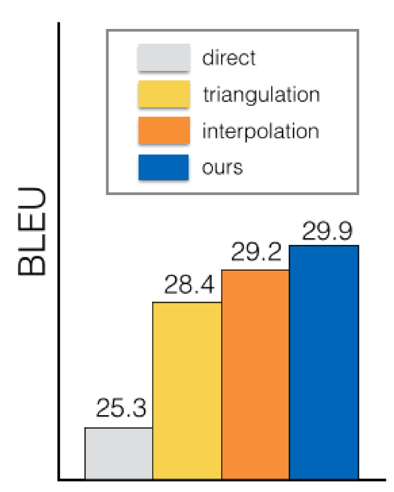
trabajan	travaillent	travail
norm. c^{sup}	1.0	-
$\hat{w}$ (triang.)	0.01	0.36
ours (no $\hat{w}$)	0.89	-
ours	0.97	0.01

Translation probabilities produced by our method better approximate the held out dev set distributions, compared to triangulation ($\hat{w}$). The dev set distributions were obtained by deriving c^{sup} from an aligned French-Spanish corpus (50k sentences) and normalizing.

Spanish-English-French

- **Direct**: 4k Spanish-French; **Triangulation**: 50k Spanish-English, 50k English-French

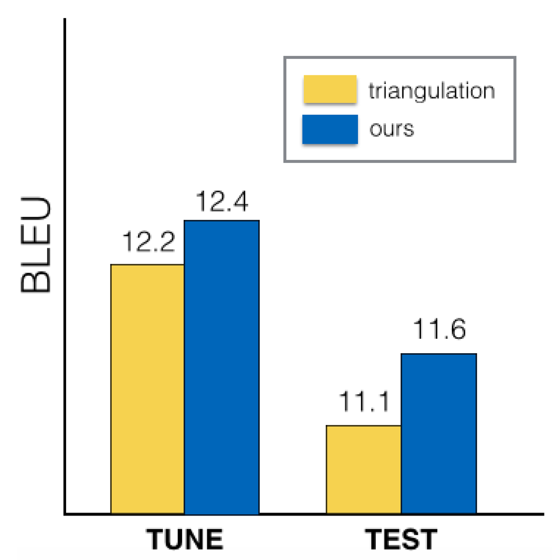




Left: Our method learns distributions closer to the dev set; **Right**: Our lexical weights improve upon interpolation by +0.7 Bleu

Malagasy-English-French

- **Direct**: 1.1k Malagasy-French dictionary entries; **Triangulation**: 100k Malagasy-English, 50k English-French



Supervision with few dictionary entries improves upon triangulation by +0.5 Bleu