Cross-lingual alignment transfer: a chicken-and-egg story?

Lauriane Aufrant[1,2], *Guillaume Wisniewski*[1], *François Yvon*[1]

(1) LIMSI-CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, (2) DGA

{lauriane.aufrant, guillaume.wisniewski, francois.yvon}@limsi.fr

Overview

- While many cross-lingual transfer techniques rely on the availability of word-aligned parallel corpora, **reliable word alignments can only be computed for large-scale parallel corpora**, a situation that is unlikely to happen for actual under-resourced languages.
- In this work, we consider transfer methods to improve the quality of word alignments for parallel corpora of very small size, eg. a few hundred sentences. We draw a typology of realistic scenarios for **cross-lingual alignment transfer** and address one of them with several baseline methods.
- From evaluation with both intrinsic and extrinsic metrics, we show that even straightforward methods can prove useful, and that **language similarities** can be successfully leveraged. We also assess that direct transfer is better done **in the domain of the bridge language**.

General transfer methods

In data space

Generate artificial target data and annotations to train target models.

↪ Direct transfer, delexicalization, annotation projection...

In parameter space

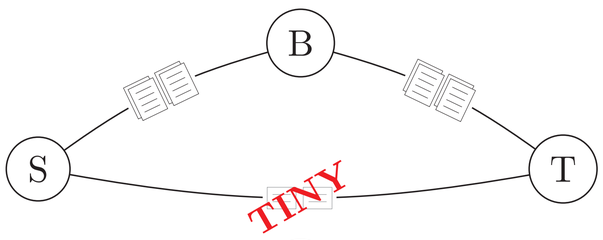
Use source model parameters to build target models.

↪ Mixture model, parameter sharing, priors...

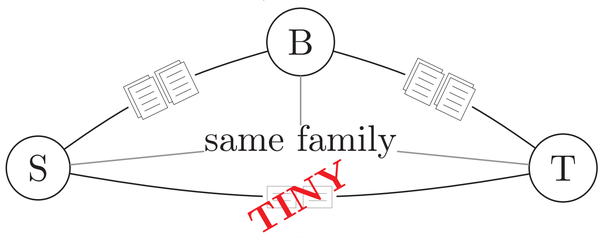
Alignment transfer: scenarios of interest

We investigate improving word alignment between two languages **S** (source) and **T** (target), using a bridge language **B**.
Depending on the availability of parallel data and on similarities between languages, we identify **5 scenarios of interest**.

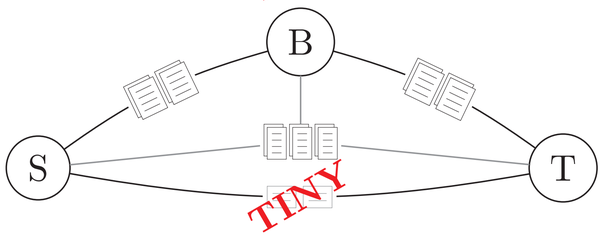
Bridge
↪ **B** = English



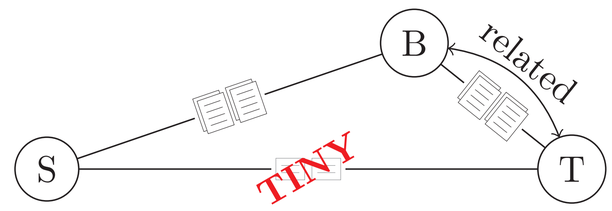
Related
↪ **S**/**B**/**T** = fr/it/ro



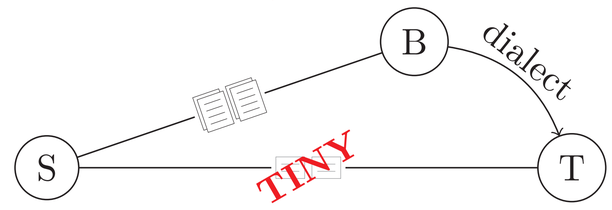
Multiparallel
↪ Europarl corpus
↪ MultiUN corpus



Directed Bridge
↪ **S**/**B**/**T** = en/ru/uk
↪ **S**/**B**/**T** = en/zh/ja



Dialect
↪ **B**/**T** = MS Arabic and dialects



Notes

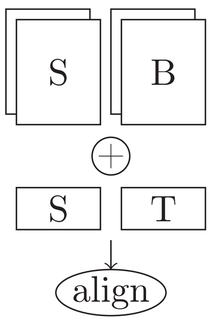
- Annotation projection is only applicable in the Multiparallel scenario.
- Delexicalized transfer raises chicken-and-egg issues.
- Two instances of Directed Bridge provide a baseline for Related.
- The method choice also depends on the task needing S-T alignments.

Baseline methods for the Directed Bridge scenario

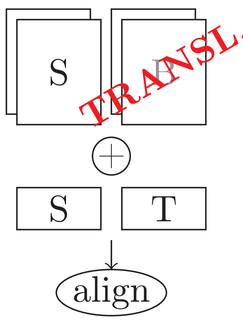
Methods for transfer in data space

The test data in pair S-T is **aligned together** with large parallel data in pair S-B.

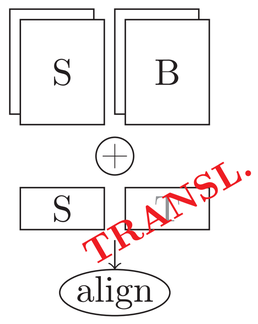
Cat-B



Tr-B



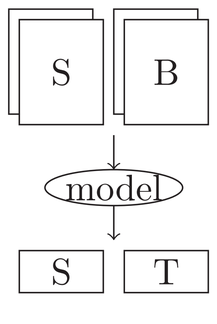
B-Tr



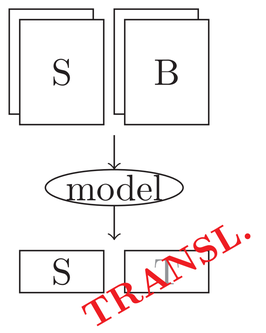
Methods for transfer in parameter space

The large S-B parallel data is used to **build an alignment model**, applied in a second step on S-T test data.

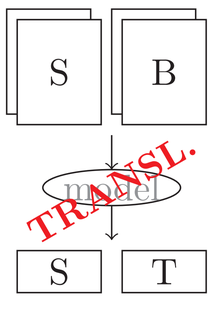
B



Glosses-B



Param-B



For each method, we investigate further improvements with word-for-word translations of parts of the data or the model.

Experiments and analysis

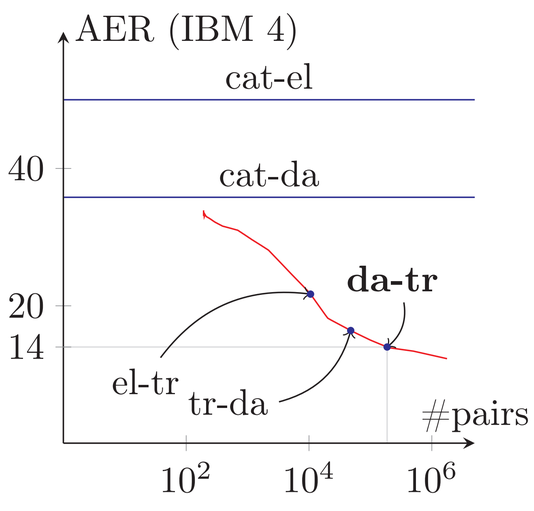
Data

- **Train:** Europarl (English, Swedish, Greek, Danish)
- **Test:** 192 English-Swedish sentence pairs, manually word-aligned [Holmqvist and Ahrenberg, 2011]

Evaluation

- **Intrinsic:** Alignment Error Rate
- **Extrinsic:** PoS accuracy of a cross-lingual PoS tagger [Wisniewski et al., 2014]

		Swedish only	Danish data			Greek data			Danish parameters		
		baseline	cat-sv	cat-da	tr-da	da-tr	cat-el	tr-el	el-tr	da	glosses-da	param-da
AER	IBM 1	53.9	**26.5**	57.0	31.1	**29.6**	74.3	**35.9**	37.4	66.0	**28.3**	33.3
	HMM	35.3	**15.3**	41.9	20.5	**16.8**	58.3	26.9	**26.4**	46.7	**16.4**	25.8
	IBM 4	33.9	**12.3**	35.8	16.4	**14.0**	50.0	**20.6**	21.7	49.1	**14.8**	24.3
PoS	IBM 1	68.7	**73.3**	58.7	73.8	**74.0**	47.4	**71.9**	71.5	67.0	**72.2**	71.1
	HMM	69.9	**73.8**	71.9	73.5	**73.6**	66.6	**73.4**	71.9	69.5	**73.4**	72.4
	IBM 4	73.0	**74.7**	74.0	73.9	**74.9**	72.0	73.4	**73.5**	66.7	**73.6**	72.0



Language similarities are leveraged

- Danish is **closely related** to Swedish, Greek is not; and Danish as a bridge clearly outperforms Greek.
- Transfer through unrelated languages is still useful.

Even straightforward methods are effective

- AER: up to **59**% relative error reduction.
- Cross-lingual task: up to **5.3**% PoS accuracy absolute improvement.

High return ratio for cross-lingual knowledge extraction

- Knowledge equivalent to one English-Swedish pair can be extracted from **five** English-Danish pairs.

A first step towards more general conclusions

- Model application with noisy proxies is better done in the **original domain** of the model.
- Even a very small piece of target data contains valuable knowledge.

NAACL'16 Workshop on Multilingual and Cross-lingual Methods in NLP