Latent Domain Word Alignment for Heterogeneous Corpora

Hoang Cuong Khalil Sima'an

Institute for Logic, Language and Computation, University of Amsterdam

Question

Heterogeneous Corpora, **Word Alignment**: What's the **link**?

Motivation

SMT with mix-of-domains haystack (a.k.a., heterogeneous data)

- We have Big DATA to train SMT systems
  - Europarl
  - United Nations
  - Common Crawl
  - News Commentary, etc.
- Wait ...
  - Data come from very different domains.
  - **How does this affect the alignment accuracy?**

  See Bach et al. (2008) and Gao et al. (2011) for reference in the literature.

Example

There are **too many** possible translations for words

- maestra → **master** (**computer**);
- maestra → **teacher** (**education**);
- maestra → **dean** (**education**);
- maestra → **crack** (**other**),
- maestra → **...** (**other**).

See Cuong and Sima'an (2014) for reference in the literature.

Our Hypothesis

Word alignment should involve **latent concepts** representing domains of data.

- IBM and HMM alignment models use context-*insensitive* conditional probabilities.
- In heterogeneous corpora the estimates of these probabilities will be **aggregated** over different domains.
- Integrating latent domain concepts into alignment models will overcome this challenge!

Contribution

A learning framework of exploiting the **contrast** between the alignment statistics **in a handful of seed samples** from different domains.

Baseline - HMM Model

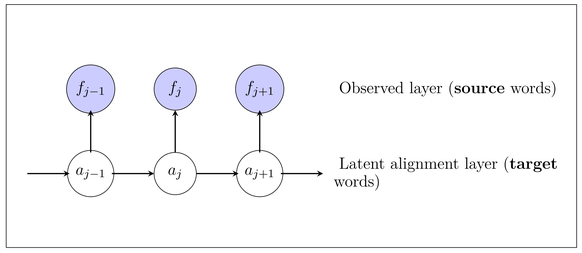

Figure: HMM alignment model.

Latent domain HMM model

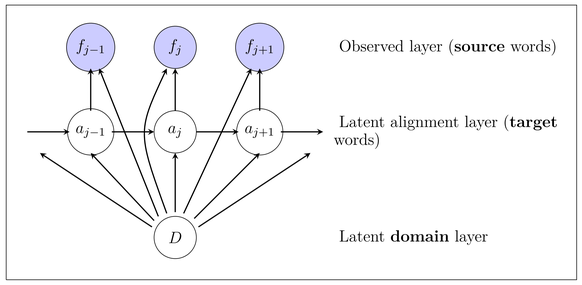

Figure: Latent domain HMM alignment model.

Learning & Decoding Algorithm

Learning

- **EM** with **Partial Supervision**
  - **Number of Domains**: The values of $D \in [1..(N+1)]$ depends on the N available seed samples plus the so-called "out-domain".
  - **Parameter Constraints**: We keep the domain prior parameters fixed for all sentence pairs that belong to seed samples.

Decoding

- New search problem (NP-hard, unfortunately):

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}} \sum_D P(\mathbf{f},\ \mathbf{a}|\ \mathbf{e},\ D) P(\mathbf{e}|\ D) P(D).$$

- We search for its lower bound instead:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}} \prod_D P(\mathbf{f},\ \mathbf{a}|\ \mathbf{e},\ D)^{P(\mathbf{e}|\ D)P(D)}$$

Result

Model	Prior	Prec.↑	Rec.↑	AER↓
	1 Million			
Baseline	-	66.95	61.29	36.00
Latent	Pharmacy (100K)	67.85	61.72	35.36
	Legal (100K)	67.57	62.29	35.17
	Hardware (100K)	69.41	63.58	33.63
	ALL (300K)	**69.64**	**63.30**	**33.68**
	2 Million			
Baseline	-	68.34	61.58	35.22
Latent	Pharmacy (100K)	68.85	62.58	34.43
	Legal (100K)	69.98	64.01	33.13
	Hardware (100K)	69.45	63.23	33.81
	ALL (300K)	**71.51**	**63.87**	**32.53**
	4 Million			
Baseline	-	69.37	64.30	33.26
Latent	Pharmacy (100K)	69.69	62.80	33.94
	Legal (100K)	70.51	63.94	32.93
	Hardware (100K)	71.75	64.44	32.10
	ALL (300K)	**72.16**	**64.30**	**31.99**

Conclusion

- Word alignment should involve latent concepts representing domains of data
- We present the benefits from learning latent domain alignment models with partial supervision.

Challenge

- Can we learn the latent domain alignment models **from scratch**?

References

[1] Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In Proceedings of COLING.

[2] Nguyen Bach, Qin Gao, and Stephan Vogel. 2008. Im- proving word alignment with language model based confidence scores. In Proceedings of the Third Workshop on Statistical Machine Translation.

[3] Qin Gao, Will Lewis, Chris Quirk, and Mei-Yuh Hwang. 2011. Incremental training and intentional over-fitting of word alignment. In Proceedings of MT Summit XIII.

[4] Hoang Cuong, Khalil Sima'an. 2014. Latent domain phrase-based models for adaptation. In Proceedings of EMNLP.

Acknowledgements

The first author is supported by the EXPERT Initial Training Network (ITN) of the European Union's Seventh Framework Programme. The second author is supported by VICI grant nr. 277-89-002 from the Netherlands Organization for Scientific Research (NWO).