Motivation

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

- We have Big DATA to train SMT systems
  - Europe:
  - 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 \rightarrow master (computer);
- maestra \rightarrow teacher (education);
- maestra \rightarrow dean (education);
- maestra \rightarrow crack (other);
- maestra \rightarrow ... (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!