A Bayesian Bilingual Segmentation Model for Transliteration

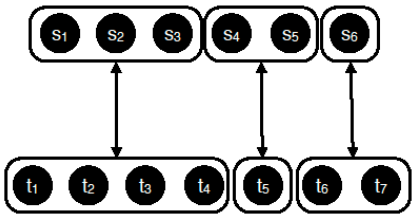
Andrew Finch and Eiichiro Sumita
Language Translation Group
National Institute of Communication Technology, Kyoto, Japan
andrew.finch@nict.go.jp, eiichiro.sumita@nict.go.jp

Overview

An unsupervised co-segmentation process for transliteration that produces a compact model with few parameters without over-fitting the training data.

Co-segmentation

Aim: to produce a co-segmentation from two bilingual sequences of tokens.



Model Characteristics

- Doesn't over-fit the training data
- Produces a compact model
- Behaves like a cache model (the rich get richer) causing re-use of segment pairs
- Expensive to introduce long sequence-pairs into the model

Generative Process

$$p((\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k))|(\bar{\mathbf{s}}_{-k}, \bar{\mathbf{t}}_{-k})) = \frac{N((\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k)) + \alpha G_0((\bar{\mathbf{s}}_k, \bar{\mathbf{t}}_k))}{N + \alpha}$$

G_0 is a base distribution, penalizes long segment-pairs.

N() is the count of the sequence pair in the history.

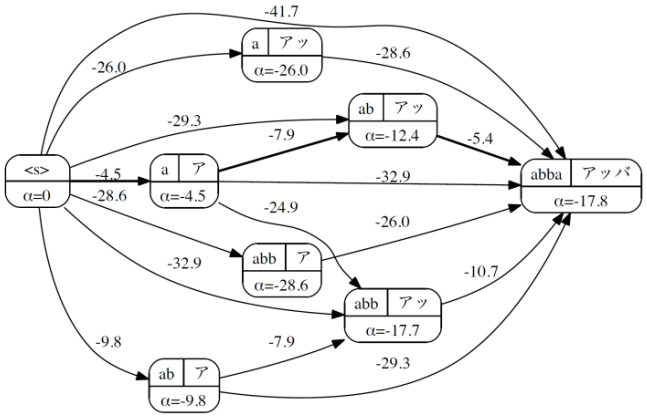
α is a parameter that controls how similar the model is to the base distribution

Training

We train the model using **blocked Gibbs sampling**. The blocks are the bi-lingual word pairs. We use forward filtering/backward sampling to efficiently sample over all co-segmentations of a bilingual word pair.

Forward filtering: Calculate the probability of the sub-graph of the co-segmentation graph back to the source node, for each node in the graph.

Backward sampling: Sample arcs backwards from sink node to source node recursively, using the probabilities calculated in the forward step.



Evaluation

Compare transliteration performance using segmentations from this approach to segmentations from GIZA++/grow-diag-final-on the NEWS2010 EN-JA data.

Phrase Extraction Model	ACC	F-score	Phrase-table Entries
GIZA++ and *grow-diag-final-and*	0.313	0.745	143382
Bayesian Segmentation	0.278	0.726	3372
Bayesian Segmentation (+agglomerated)	0.323	0.748	102507
Bayesian Segmentation (+integrated)	0.329	0.752	164258

MASTAR

NiCT National Institute of Information and Communications Technology