Bitexts

P. S. Langeslag

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA
SEIT 1737

Bitexts in the *Dictionary of Old English* Corpus (DOEC)

- Psalter glosses (13 copies; c. 70,000 words per complete bitext)
- Gospel glosses (2 copies; c. 135,000 words per complete bitext)
- Taunton fragment (4 fragmentary macaronic homilies; 1000 words)

- Some 436,200 Latin words to match c. 492,600 Old English words.
- A lot of duplication

Readily Available Old English–Modern English Bitexts

- Ælfric's *Lives of Saints* on Wikisource

Statistical Machine Translation vs Encoder–Decoder Model

Statistical Machine Translation

- Relies on explicit word alignment

Encoder–Decoder Model

- No need for explicit word alignment

Continued relevance of word alignment:

- Semantic mining of small-resource languages

Cf. Li, "Word Alignment in the Era of Deep Learning: A Tutorial."

Discriminative vs Generative Models 1/2

Discriminative

- Supervised: require a modest amount of labelled training data
- Base a discriminant function for $P(y|x)$ directly on training data
- Focus on the alignment prediction
- Require careful feature selection (co-occurrence, prior predictions, token length, generative predictions, etc.)
- Learned models don't generalize well across domains or languages

Generative

- Unsupervised
- Infer $P(y)$ and $P(x|y)$ from training data, then calculate $P(y|x)$ using the Bayes Theorem
- Focus on uncovering the hidden alignment behind observations, prediction a byproduct
- Strong independence assumptions required to avoid too complex a model
- e.g. IBM alignment models

Discriminative vs Generative Models 2/2

> “In order to make discriminative alignment competitive with unsupervised generative approaches, one needs to show that language-independent features can be used with high confidence on various domains.” (Tiedemann 99)

IBM Alignment Models 1–2

Model 1

- Assumes a uniform distribution across possible alignment positions, ignoring position/word order
- Convex function, so the local optimum is always the global optimum
- Of little use except to establish the initial parameters for the other, non-convex models

Model 2

- Adds position (reordering) based on absolute token positions
- Two steps: (1) lexical translation; (2) reordering (based on target token order)
- Not convex, thus relies on model 1 for initial parameters

IBM Alignment Model 3

- Adds support to align source tokens to multiple target tokens or none, as well as to insert source NULL tokens
- Four steps: (1) fertility; (2) NULL insertion; (3) lexical translation; (4) distortion

fertility	number of target tokens a source token can generate
distortion	predicts target position based on source position

IBM Alignment Models 4–6

Model 4

- Distorts on the basis of relative rather than absolute position
  - Based on the placement of target tokens generated by preceding words
  - Accounts for groups of words (**cepts**) moving together
- Adds an (implicit) reliance on word classes
- The standard in statistical alignment

Model 5

Like model 4, but solves undefined placements at additional training cost.

Model 6

Like model 4, but incorporating a Hidden Markov Model (see next slide).

First-Order Hidden Markov Model

Assumes that we can predict the next link solely on the basis of the current link.

IBM Models in Sum

- Purely statistical and thus language-agnostic
- Models 4–6 implicitly accommodate POS information
- They function as a pipeline, each relying on the previous model
- A Hidden Markov Model is often inserted between models 2 and 3 in the pipeline

Using Additional Data

- Bilingual dictionaries

Selected Drawbacks to Generative Alignment Models

- Asymmetrical, direction-dependent alignment
- Work best on similar languages

`fast_align`

- Based on IBM model 2
- Written in C++, available in Python through `systran-align`

Giza++

- Based on IBM models 1–5 plus a Hidden Markov Model
- Trains word classes using a maximum-likelihood criterion (`mkcls`)

Now maintained as part of https://github.com/moses-smt/.

Bibliography

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Computational Linguistics* 19, no. 2 (1993): 263–311.

Li, Bryan. "Word Alignment in the Era of Deep Learning: A Tutorial." arXiv, 2022. https://doi.org/10.48550/ARXIV.2212.00138.

Tiedemann, Jörg. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies 14. Springer, 2011.