Bilingual Lexicon Induction using Small Quantities of Sentence-Aligned Phonemic Transcriptions

Oliver Adams[1], Graham Neubig[2], Trevor Cohn[1], Steven Bird[1]

[1]The University of Melbourne
[2]Nara Institute of Science and Technology

Background

- The world's languages are dying out.
- There is a movement to speed up data collection using more ad hoc approaches.

Background

- A modest quantity is bilingual.
- Bilingual lexicons are an important part of language documentation and they are valuable in downstream NLP.

Question

v I 6 v I s @ n n I C t v a s p a s i: 6 t

We do not know what is happening

Given a tiny quantity of data of this sort, how well can we learn bilingual lexical entries by training translation models and extracting high confidence entries?

Challenges

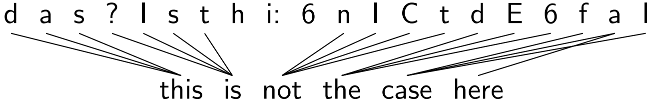
1. Limited data
2. No word segmentation
3. Erroneous phoneme recognition

Models

1. Traditional word alignment with GIZA++
2. Model 3P: an adapted IBM Model 3
3. Unsupervised word segmentation (UWS) followed by word alignment
4. Joint segmentation and alignment using a Bayesian inversion transduction grammar (ITG) model

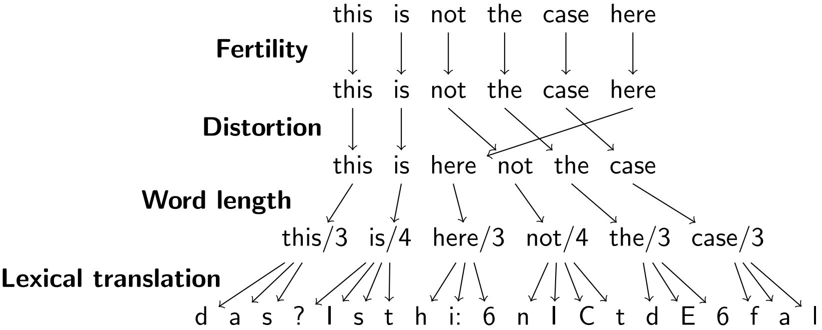
1. GIZA++

- Traditional word alignment baseline using the IBM models.
- Learns Lexical translation probabilities that relate source tokens to the target tokens.
- But our source side tokens are phonemes and can't be meaningfully translated into English words.



2. Model 3P

- Introduced by Stahlberg et. al (IEEE SLT 2012) as implemented in PISA.[1]
- Extends IBM Model 3 to include a word length parameter.
- Generation of phonemes is conditioned on an English word and a phoneme position in a target word.



[1] https://code.google.com/p/pisa/

3. UWS GIZA++

1. Break phoneme sequence into chunks with pgibbs.[2]
2. Perform alignment with GIZA++

das ?Is t hi:6 nIC t dE6 fal

this is not the case here

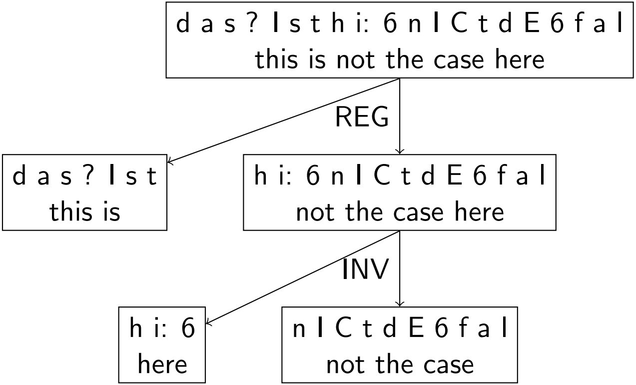
[2] https://github.com/neubig/pgibbs

4. Bayes ITG

- Uses Gibbs sampling to derive inversion transduction grammar trees (as implemented in pialign[3]).
- Each tree describes how the German relates to English.
- Hierarchical nature permits phrases of varying granularities.

[3] https://github.com/neubig/pialign

Joint segmentation and alignment with a Bayesian ITG model



Data

1. Start with German–English data from Europarl.
2. Remove punctuation.
3. Convert German to a sequence of phonemes using a text-to-speech system (MARY[4]).
4. Remove stress markers and syllable boundaries that ASR systems can't reasonably capture.
5. Limit training sentences to be fewer than 100 phonemes.
6. Break into 1k, 2k, 5k and 10k sentence training sets.

[4] http://mary.dfki.de/

Lexicon induction approach

1. Train the translation models.
2. Filter out entries in the phrase tables that include only one phoneme.
3. Sort entries by their joint probability.
4. Filter for the top 5 entries of an English word given a phonemic entry, and vice versa.
5. Consider the top 500 entries for evaluation.

Annotation

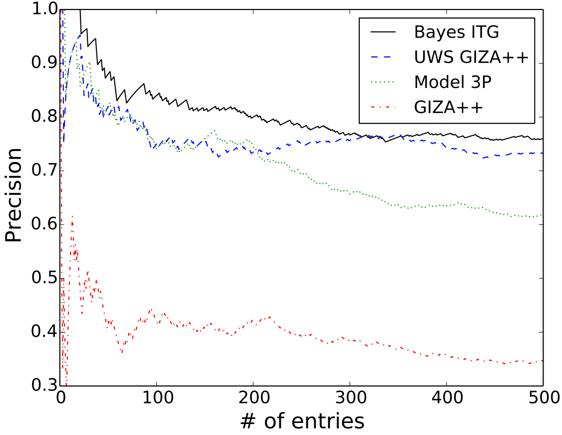
Entries from the bilingual lexicons were merged, shuffled and given to a German for annotation. They were marking them as *correct*, *incorrect*, or *ambiguous*:

- *Correct* entries are found in a German–English dictionary.
- *Incorrect* entries are deemed “clearly incorrect” by the annotator. Examples:
  - *tsu:?aIn*⇔*the* (“zu ein”)
  - *b@dINUN*⇔*be* (“Bedingung”)
- *Ambiguous* entries are neither strictly correct nor incorrect. Examples:
  - *nvi:6*⇔*we* (“wir”)
  - *nICt*⇔*does not* (“nicht”)

Evaluation

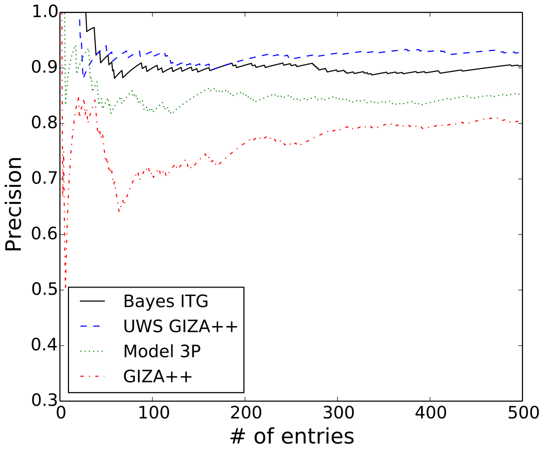
Precision at k entries

Over 10k sentences, with only strictly correct entries as valid.



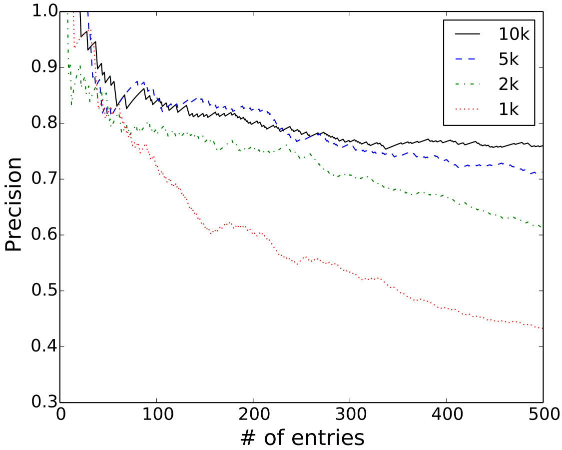
Precision at k entries

Over 10k sentences, considering ambiguous entries as valid.



Precision at k entries

Bayes ITG, with only strictly correct entries helping precision.



Monolingual lexical entries

- Many entries consisted of correctly segmented phonemes misaligned to English (*Bedingung* as in *b@dINUN*⇔*be*).
- But monolingual entries are useful in their own right.

Monolingual lexical entry performance

The accuracy of the segmentation of phonemic lexical entries judged incorrect and ambiguous.

Method	Sents	Incorrect %	Correct seg. %
Bayes ITG	1k	26.2	52.7
Bayes ITG	2k	16.6	60.2
Bayes ITG	5k	13.4	62.7
Bayes ITG	10k	9.6	62.5
UWS GIZA++	10k	7.2	38.9
GIZA++	10k	19.4	15.5
Model 3P	10k	14.6	46.6

Word segmentation performance

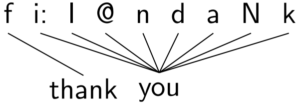
Note that UWS GIZA++ is still informed by the English. Without the English most of the entries aren't words.

Token	Occurrences
?	13,096
@	8,587
n	8,138
t	6,422
@n	6,300
d	5,929
s	3,226
6	3,136
f	3,099
di:	2,913

Qualitative observations of the entries

Observations: segmentation quality

- Bayes ITG approach tends to grab words and multi-word expressions cleanly, even if misaligned.
- Model 3P and UWS GIZA++ tend to have more off-by-one errors and alignments such as this:



Observations: segmentation granularity

- M3P tended to bias towards shorter units
- Bayes ITG was more likely to use longer phrases:

$$tvo6d@n \Leftrightarrow been$$

Observations: UWS GIZA++ nuances

UWS GIZA++ errors were more distinct:

- *t?*$\Leftrightarrow$*is*
- *n?*$\Leftrightarrow$*to*
- *n?*$\Leftrightarrow$*of*

Likely a result of the pipelined nature.

Conclusions

- Lexical entries can be learnt with decent precision even with very small quantities of data.
- These are applicable when small quantities of reliable phonemic transcriptions are available.
- Future work should consider real data with errors from acoustic models, as that is a significant bottleneck we did not address.