Fast and Accurate Sentence Alignment of Bilingual Corpora

Robert C. Moore

Presented by
Kathrin Adlung

Overview

Introduction

Description of the Algorithm

Sentence-Length-Based Alignment

Search Issues

Word Translation Model

Word-Correspondence-Based Alignment

Results

Conclusion

Introduction

- Machine learning → machine translation
- Should be fast, highly accurate and require no special knowledge
- First approach: modelling the relationship between the lengths of sentences that are mutual translations

Introduction

- Brown: similar to first approach
- Gale and Church: similar to first approach
- Chen: optimizing word-translation probabilities
- Wu: adapted G.a.C. to Chinese with lexical cues to improve alignment accuracy
- Melamed: based on word correspondences
- Simard and Plamondon: two-pass approach

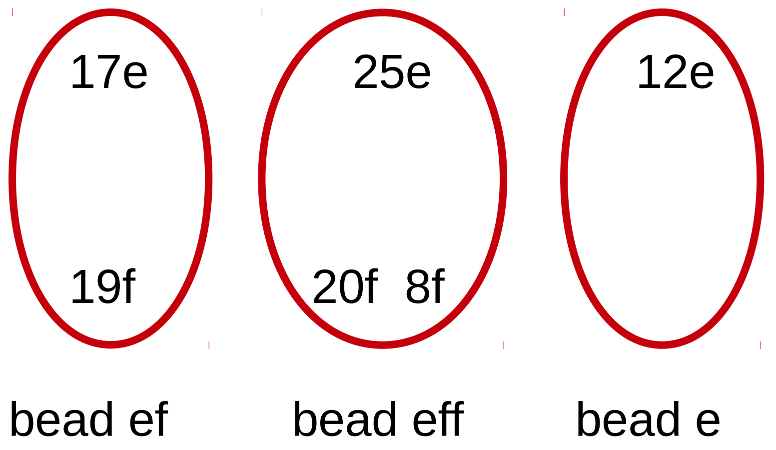
→ require particular knowledge about the corpus or the language involved

Description of the Algortihm

- Hybrid method
- 3 Step process
  1. Align the corpus
  2. Train a modified version of IBM Translation Model 1
  3. Realign the corpus
- Requires no externally supplied lexicon

Sentence-Length-Based Alignment

Major anchorpoint	Major anchorpoint
S1 S2 S2	S1 S2 S2
Minor anchorpoint	Minor anchorpoint
S1 S2 S3 S4 S5	S1 S2 S3 S4 S5
S6 S7 S8	S6 S7 S8
Minor anchorpoint	Minor anchorpoint
...	...
Minor anchorpoint	Minor anchorpoint
...	...
Major anchorpoint	Major anchorpoint
...	...
Minor anchorpoint	Minor anchorpoint
...	...



Sentence-Length-Based Alignment

Gaussian distribution

$$P(l_t|l_s) = \alpha\ exp(-((log(l_t/l_s) - \mu)^t/2\sigma^2))$$

$$r = \log(\ell_f/\ell_e)$$

$$\Pr(\ell_f|\ell_e) = \alpha \exp[-(r - \mu)^2/(2\sigma^2)]$$

Poisson distribution

$$P(l_t|l_s) = exp(-l_s r)(l_s r)^{l_t}/(l_t!)$$

Sentence-Length-Based Alignment

- Differences between the models:
  - (Brown) estimates marginal distribution of sentence lengths using raw relative frequencies
  - (Moore) using raw relative frequencies to estimate the probability of every sentence length
  - Model was insensitive to the exact values of the probabilities assigned to the bead types
  - No paragraph boundaries
- Intention is not to improve its accuracy
  - Faster to use in practice

Search Issues

- Standard approach for alignment
  - Dynamic programming (DP)
    - Infeasible for large corpora

→ search must be pruned

  - Novel approach

Search Issues

- True points of correspondence should all be close to proportionately the same distance from the beginning of each text
  - The set of points similar as forming a matrix
- Pruned DP does exhaustive search
  - Only narrow fixed-width band around the main diagonal

Search Issues

- How do we know this is the case?
  - Heuristic algorithm
    - Find best alignment within the band
    - Otherwise widen the band

- Never seen that the heuristic committed a search error

- Find all high-probability 1 -to- 1 beads
  - To train a word translation mode
    - Forward-backward probability computation

Word Translation Model

- Modified IBM Translation Model 1
- How target language sentence is generated from source language sentence

$$P(t|s) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} tr(t_j|s_i)$$

1. A length m is selected for t
2. For each word position in t a generated word in s is selected
3. For each pair of a position in t and its generating word in s a target language word is chosen

Word Translation Model

- Modifications
  - Translation probabilities for rare words are omitted
  - Words get a minimum number of occurences
    - Words with fewer occurences will be mapped into a single token
  - Word-translation pairs that dont reach a specific fractional count are not added.

Word-Correspondence-Based Alignment

- Final model: modified to use IBM Model 1 in addition to the initial model.
- Model assumes bead types and sentence lengths are generated according to the same probability distribution.
- 1 -to- 0, 0 -to- 1 bead types
  - Each word is generated independently
- 1 -to- 1, 1 -to- 2, 2 -to- 1 bead types
  - Words of source language as 1 -to- 0, 0 -to- 1
  - Words of target language depending on word of source language

Word-Correspondence-Based Alignment

- In applying Model 1 omit uniform distribution of target sentence length
- Combined model:

$$P(s,t) = \frac{P_{1-1}(l,m)}{(l+1)^m}\left(\prod_{j=1}^{m}\sum_{i=0}^{l} tr(t_j|s_i)\right)\left(\prod_{i=1}^{l} f_u(s_i)\right)$$

s = source sentence
l = length of s
t = target sentence
m = lengrh of t
$P_{1-1}(l,m)$ = probability of lenth l aligning 1 -to- 1 with a sentence of length m

Word-Correspondence-Based Alignment

- Hybrid alignment model incorporating IBM Model 1
- Limited search
- Final alignment search faster than initial alignment search

Results

Table 1. Results for Manual 1 data

Alignment Method	Probability Threshold	Number Right	Number Wrong	Number Omitted	Precision Error	Recall Error
Hand-Aligned	NA	9842	1	6	0.010%	0.061%
Length Only	0.5	9832	28	16	0.284%	0.162%
Length+Words	0.5	9846	5	2	0.051%	0.020%
Length+Words	0.9	9839	3	9	0.030%	0.091%

Table 2. Results for Manual 2 data

Alignment Method	Probability Threshold	Number Right	Number Wrong	Number Omitted	Precision Error	Recall Error
Hand-Aligned	NA	17276	5	99	0.029%	0.570%
Length Only	0.5	17304	18	71	0.104%	0.409%
Length+Words	0.5	17361	2	14	0.012%	0.081%
Length+Words	0.9	17316	1	59	0.006%	0.340%

Results

Table 3. Results for Manual 1 data with deletions

Sentences Deleted	Alignment Method	Number Right	Number Wrong	Number Omitted	Precision Error	Recall Error
0	Length Only	9832	28	16	0.284%	0.162%
50	Length Only	9761	30	39	0.306%	0.398%
100	Length Only	9677	30	73	0.309%	0.749%
300	Length Only	9368	52	187	0.552%	1.967%
0	Length+Words	9846	5	2	0.051%	0.020%
50	Length+Words	9796	6	4	0.061%	0.041%
100	Length+Words	9747	5	3	0.051%	0.031%
300	Length+Words	9550	4	5	0.042%	0.052%

Results

Table 4. Alignment time (in seconds) for deletion experiments

Sentences Deleted	First Pass Iterations	Length Align Time	Model 1 Train Time	Length+Words Align Time	total Total
0	1	161	131	155	447
50	3	686	133	195	1013
100	5	1884	128	281	2293
300	7	4360	125	555	5040

Conclusion

1. Modification of Brown et al.'s sentence-length-based model to use Poisson distributions, rather than Gaussians, so that no hidden parameters need to be iteratively re-estimated.
2. A novel iterative-widening search method for alignment problems, based on detecting when the current best alignment comes near the edge of the search band, which eliminates the need for anchor points.
3. Modification of IBM Translation Model 1, eliminating rare words and low probability translations to reduce the size of the model by 90% or more.
4. Use of the probabilities computed by a relatively cheap initial model (the sentence-length-based model) to dramatically reduce the search space explored by a second more accurate, but more expensive model (the word-correspondence-based model). While this idea has been used in such fields as speech-recognition and parsing, it seems not to have been used before in bilingual alignment.