Approaches to Machine Translation: Rule-based, Statistical and Hybrid

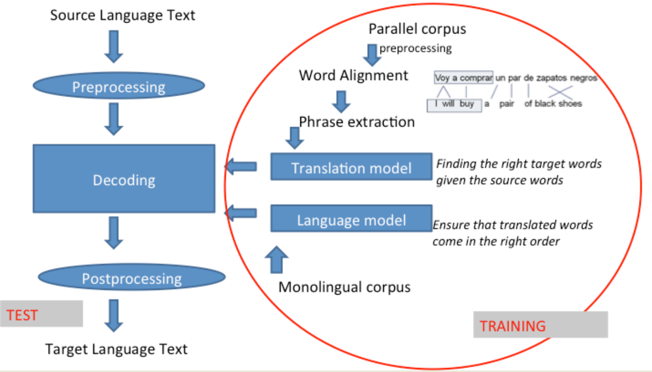
Alignment - IBM Model 2 and HMM (III)

Lluís Formiga

July 11, 2014

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

A picture is worth a million equations



Outline

- ~~Noisy Channel Model~~
- ~~Lexical translation~~
- ~~Word Alignment~~
- ~~Expectation Maximization (EM) Algorithm~~
- IBM Models 1--5
  - ~~IBM Model 1: lexical translation~~
  - IBM Model 2: alignment model
  - IBM Model 3: fertility
  - IBM Model 4: relative alignment model
  - IBM Model 5: deficiency
- HMM Models: dependent alignment model
- Problems of Word Alignment
- Quality of Word Alignment

IBM Model 2

Absolute Alignment Model

Basic statistics:

- IBM Model 1 captures $p(f|e)$

Translations of *mesa*	$p(f\|e)$
table	0.3771
round	0.1476
panel	0.1344
round-table	0.0452
petitioners	0.0282
bureau	0.0229
officers	0.0190
Committee	0.0169
Round	0.0153
roundtable	0.0124

Also we can collect advanced statistics:

- IBM Model 2 captures $q(j|i, l_f, l_e)$

j	i	l_f	l_e	$q(j\|i, l_e, l_f)$
1	1	5	7	0.27
1	2	5	7	0.14
⋮	⋮	5	7	0.07
5	7	5	7	1e-14
1	1	5	8	0.32
1	2	5	8	0.18
⋮	⋮	8	6	0.13
5	8	5	8	1e-19
⋮	⋮	⋮	⋮	⋮
1	1	6	8	0.30
1	2	6	8	0.12
⋮	⋮	6	8	0.17
6	8	6	8	1e-10

IBM Model 2

Definition

- Modeling alignment with an alignment probability distribution
- Translating source word at position j to translated word at position i:

$$q(i|j, l_e, l_f)$$

- Source position conditioned on target and lengths
- Putting everything together:

$$p(e, a|f) = \epsilon \prod_{j=1}^{l_f} t(e_i|f_j) \cdot q(i|j, l_e, l_f)$$

- EM training of this model works the same way as IBM Model 1

Generative Model

What is Statistical Machine Translation?

- Series of chained decisions
- Happening with certain probability
- Probability of the product
  - The translation can be produced by different ways
  - We sum up the probabilities of all the ways

e.g: IBM Models 1 and 2

$$
\begin{aligned}
P(e|f) &= \sum_a P(e,a|f) = p(f,a|e) \cdot p(e) \\
&= \sum_a \frac{\epsilon}{(l_e+1)^{l_f}} \prod_{j=1}^{l_f} t(f_j|e_{a(j)}) \cdot [(j|a(j), l_e, l_f] \cdot p(e)
\end{aligned}
$$

IBM Model 2

Partially Observed Data

Require: set of sentence pairs $(\mathbf{e}_s, \mathbf{f}_s)$
Ensure: translation prob. $t(f|e)$ for all foreign words f and end words e
Ensure: alignment prob. $q(j|i, l_e, l_f)$ for all foreign positions j and end positions i given lengths l_e and l_f

{**initialize** $t(f|e)$ **and** $q(j|i, l_e, l_f)$ **uniformly or from other training** }
repeat
  $\forall_{e_i \in e} \forall_{f_j \in f} : \text{count}(e_i|f_j) = 0$
  $\forall_{f_j \in f} : \text{total}(f_j) = 0$
  {**compute normalization** }
  for all sentence pairs (e_s, f_s) **do**
    for all words $f_j \in f_s$ **do**
      $\text{total}_s(f_j) = 0$
      for all words $e_i \in e_s$ **do**
        $\text{total}_s^a(j) \mathrel{+}= t(f_j|e_i) \cdot q(j|i, l_e, l_f)$
      end for
    end for
    {**collect counts** }
    for all words $f_j \in f_s$ **do**
      for all words $e_i \in e_s$ **do**
        $\text{count}(f_j, e_i) \mathrel{+}= \frac{t(f_j|e_i) \cdot q(j|i, l_e, l_f)}{\text{total}_s^a(f_j)}$
        $\text{total}(e_i) \mathrel{+}= \frac{t(f_j|e_i) \cdot q(j|i, l_e, l_f)}{\text{total}_s^a(f_j)}$
        $\text{count}(j, i, l_e, l_f) \mathrel{+}= \frac{t(f_j|e_i) \cdot q(j|i, l_e, l_f)}{\text{total}_s^a(f_j)}$
        $\text{total}(i, l_e, l_f) \mathrel{+}= \frac{t(f_j|e_i) \cdot q(j|i, l_e, l_f)}{\text{total}_s^a(f_j)}$
      end for
    end for
  end for
  {**estimate probabilities** }
  for all words $f_j \in f$ **do**
    for all words $e_i \in e$ **do**
      $t(f_j|e_i) = \frac{\text{count}(f_j, e_i)}{\text{total}(e_i)}$
      $q(j|i, l_e, l_f) = \frac{\text{count}(j, i, l_e, l_f)}{\text{total}(i, l_e, l_f)}$
    end for
  end for
until convergence

Why are these algorithms so simple?

- Each word and alignment link are generated separately; there are no dependencies between alignment links at all
- ⋆ The cost of easy inference here is an overly simplistic model

Some drawbacks of word based alignments

- All reorderings have the same probability
- Alignments are independent
- No notion of multiword alignments
- Alignments are asymmetric
- No morphology
- No syntax

Models 1 & 2

Strengths & Weaknesses

Some drawbacks of word based alignments

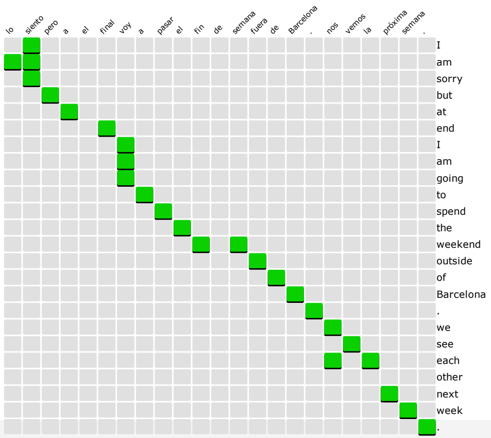
- ~~All reorderings have the same probability~~ MODEL 2
- Alignments are independent
- No notion of multiword alignments
- Alignments are asymmetric
- No morphology
- No syntax

HMM Models

Vogel, Ney, & Tillmann ('96)

- Motivation
  - Strong localization effect in aligning the words in parallel texts (∀ language pairs ∈ Indoeuropean)
  - Words not distributed arbitrarily over the sentence positions forming clusters.
  - Alignments mostly preserve local neighborhood
  - Most cases with stronger restriction:

→ the difference in the position index is smaller than 3.



HMM Models

Model 2 used the absolute positions of words

$$p(a|e,m) = \prod_{i=1}^{m} q(a_i = j|i, l_f, l_e)$$

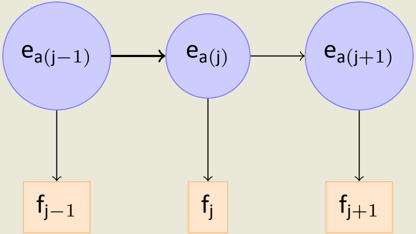
A better idea: relative positioning using position differences

$$p(a|e,m) = \prod_{i=1}^{m} q(a_i|a_{i-1})$$

- A local shift probability

HMM Models

Main idea:



Formally:

$$p(e, a|f) = \epsilon \prod_{j=1}^{l_f} t(f_j|e_{a(j)}) p_a(a(j)|a(j-1), l_e)$$

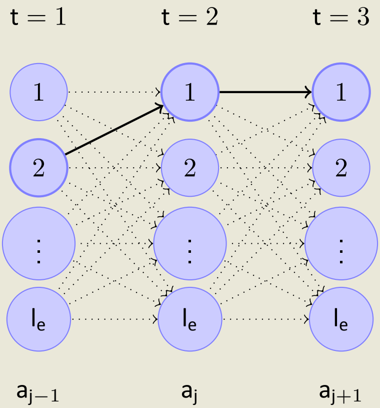
HMM Models

Alignment model is parametrized with a simple global table

- Alignment links are no longer conditionally independent!
- Inference (and EM) now require something more complicated (dynamic programming)

Shift distance	Prob
-3	0.03
-2	0.05
-1	0.12
0	0.2
1	0.3
2	0.09
3	0.08

Dynamic Programming



Models 1 & 2

Some drawbacks of word based alignments

- ~~All reorderings have the same probability~~ MODEL 2
- ~~Alignments are independent~~ HMM MODEL
- No notion of multiword alignments
- Alignments are asymmetric
- No morphology
- No syntax

Next session

- ~~Noisy Channel Model~~
- ~~Lexical translation~~
- ~~Word Alignment~~
- ~~Expectation Maximization (EM) Algorithm~~
- IBM Models 1--5
  - ~~IBM Model 1: lexical translation~~
  - ~~IBM Model 2: alignment model~~
  - IBM Model 3: fertility
  - IBM Model 4: relative alignment model
  - IBM Model 5: deficiency
- ~~HMM Models: dependent alignment model~~
- Problems of Word Alignment
- Quality of Word Alignment