Improved Word Alignments for Statistical Machine Translation

Alex Fraser
Institute for NLP
University of Stuttgart
Statistical Machine Translation (SMT)

- Build a model $P( e | f )$, the probability of the English sentence “$e$” given the French sentence “$f$”
- To translate a French sentence “$f$”, choose the English sentence “$e$” which maximizes $P( e | f )$

$$\arg\max_e P( e | f ) = \arg\max_e P( f | e ) P( e )$$

- $P( f | e )$ is the “translation model”
  - Collect statistics from word aligned parallel corpora
- $P( e )$ is the “language model”
Annotation of Minimal Translational Correspondences

- Word alignment is annotation of minimal translational correspondences
- Annotated in the context in which they occur
- Not idealized translations!

(solid blue lines annotated by a bilingual expert)

Alex Fraser
Overview

• Solving problems with previous word alignment methodologies
  – Problem 1: Measuring quality
  – Problem 2: Modeling
  – Problem 3: Utilizing new knowledge
  – Joint Work with Daniel Marcu, USC/ISI
Problem 1: Existing Metrics Do Not Track Translation Quality

- Dozens of papers report word alignment quality increases according to intrinsic metrics
- Contradiction: few of these report MT results; those that do report inconclusive gains
- This is because the two commonly used intrinsic metrics, AER and balanced F-Measure, do not correlate with MT performance!
Measuring Precision and Recall

• Start by fully linking hypothesized alignments

• Precision is the number of links in our hypothesis that are correct
  – If we hypothesize there are no links, have 100% precision

• Recall is the number of correct links we hypothesized
  – If we hypothesize all possible links, have 100% recall

• We will test metrics which formally define and combine these in different ways

Alex Fraser
Alignment Error Rate (AER)

Gold

Precision\((A, P) = \frac{|P \cap A|}{|A|} = \frac{3}{4}\) (e3,f4) wrong

Recall\((A, S) = \frac{|S \cap A|}{|S|} = \frac{2}{3}\) (e2,f3) not in hyp

Hypothesis

\[
\text{AER}(A, P, S) = 1 - \frac{|P \cap A| + |S \cap A|}{|S| + |A|} = \frac{2}{7}
\]

BLUE = sure links GREEN = possible links

Alex Fraser
Experiment

• Desideratum:
  – Keep everything constant in a set of SMT systems except the word-level alignments
    • Alignments should be realistic

• Experiment:
  – For better alignments: train on 16M, 32M, 64M words (but use only the 8M words for MT building).
  – For worse alignments: train on $2 \times 1/2$, $4 \times 1/4$, $8 \times 1/8$ of the 8M word training corpus.

• If AER is a good indicator of MT performance, $1 - \text{AER}$ and BLEU should correlate no matter how the alignments are built (union, intersection, refined)
  – Low $1 - \text{AER}$ scores should correspond to low BLEU scores
  – High $1 - \text{AER}$ scores should correspond to high BLEU scores
AER is not a good indicator of MT performance

\[ r^2 = 0.16 \]
**F_α-score**

Gold:

\[
\begin{align*}
\text{f1} & \quad \text{f2} \quad \text{f3} \quad \text{f4} \quad \text{f5} \\
\hline
\text{e1} & \quad \text{e2} \quad \text{e3} \quad \text{e4}
\end{align*}
\]

Hypothesis:

\[
\begin{align*}
\text{f1} & \quad \text{f2} \quad \text{f3} \quad \text{f4} \quad \text{f5} \\
\hline
\text{e1} & \quad \text{e2} \quad \text{e3} \quad \text{e4}
\end{align*}
\]

\[
\text{Precision}(A, S) = \frac{|S \cap A|}{|A|} = \frac{3}{4} \quad \text{(e3,f4)} \quad \text{wrong}
\]

\[
\text{Recall}(A, S) = \frac{|S \cap A|}{|S|} = \frac{3}{5} \quad \text{(e2,f3)} \quad \text{(e3,f5)} \quad \text{not in hyp}
\]

\[
\text{F}(A, S, \alpha) = \frac{1}{\alpha \text{Precision}(A, S) + \frac{1-\alpha}{\text{Recall}(A, S)}}
\]

Called F_α-score to differentiate from ambiguous term F-Measure.
$F_\alpha$-score is a good indicator of MT performance

$r^2 = 0.85 \quad \alpha = 0.4$
Discussion

• Using $F_\alpha$-score as a loss criterion will allow for development of discriminative models (later in talk)

• AER is not derived correctly from F-Measure

• For details of experiments see squib in Sept. 2007 Computational Linguistics
Problem 2: Modeling the Wrong Structure

- 1-to-N assumption
  - Multi-word “cepts” (words in one language translated as a unit) only allowed on target side. Source side limited to single word “cepts”.
- Phrase-based assumption
  - “cepts” must be consecutive words
Explicitly model three word types:

- **Head word**: provide most of conditioning for translation
  - Robust representation of multi-word cepts (for this task)
  - This is to semantics as ''syntactic head word'' is to syntax
- **Non-head word**: attached to a head word
- **Deleted source words** and **spurious target words** (NULL aligned)
### LEAF Generative Story

- Once sourcecepts are determined, exactly one target head word is generated from each source head word.
- Subsequent generation steps are then conditioned on a single target and/or source head word.
- See EMNLP 2007 paper for details.
Can score the same structure in both directions

Math in one direction (please do not try to read):

\[
p(f, a|e) = \left[ \prod_{i=1}^{l} g(\chi_i | e_i) \right] \\
\left[ \prod_{i=1}^{l} \delta(\chi_i, -1) w_{-1}(\mu_i - i|\text{class}_c(e_i)) \right] \\
\left[ \prod_{i=1}^{l} \delta(\chi_i, 1) t_1(\tau_{i1} | e_i) \right] \left[ \prod_{i=1}^{l} \delta(\chi_i, 1) s(\psi_i | e_i, \gamma_i) \right] \\
\left[ s_0(\psi_0) \sum_{i=1}^{l} \psi_i \right] \left[ \prod_{k=1}^{t_0} t_0(\tau_{0k}) \right] \\
\left[ \prod_{i=1}^{l} \prod_{k=2}^{t_1} t_{>1}(\tau_{ik} | e_i, \text{class}_h(\tau_{i1})) \right] \\
\left[ \prod_{i=1}^{l} \prod_{k=1}^{D} D_{ik}(\pi_{ik}) \right]
\]
Discussion

• LEAF is a powerful model
• But, exact inference is intractable
  – We use hillclimbing search from an initial alignment
• First model of correct structure: M-to-N discontiguous
  – Head word assumption allows use of multi-word cepts
    • Decisions robustly decompose over words
    • Does not have segmentation problem of phrase alignment models: Probability of alignments of cept “the man” are closely related to probabilities for cept “man”
  – Not limited to only using 1-best prediction
Problem 3: Existing Approaches Can’t Utilize New Knowledge

• It is difficult to add new knowledge sources to generative models
  – Requires completely reengineering the generative story for each new source

• Existing unsupervised alignment techniques can not use manually annotated data
Background

• We love EM, but
  – EM often takes us to places we never imagined/wanted to go
• Bayes is always right

\[
\text{argmax } P(e \mid f) = \text{argmax } P(e) \times P(f \mid e)
\]

But in practice, this works better:

\[
\text{argmax } P(e)^{2.4} \times P(f \mid e) \times \text{length}(e)^{1.1} \times \text{KS}^{3.7} \ldots
\]
Decomposing LEAF

- Decompose each step of the LEAF generative story into a sub-model of a log-linear model
  - Add backed off forms of LEAF sub-models
  - Add heuristic sub-models (do not need to be related to generative story!)
  - Allows tuning of vector $\lambda$ which has a scalar for each sub-model controlling its contribution
Reinterpreting LEAF

- $g(e_i)$ – source word type sub-model
- $w(\mu_i)$ – source non-head linking sub-model
- $t_1(f_j | y(i))$ – head word translation sub-model
- Etc… – many more sub-models

\[
p(a, f | e) = g \times w \times t_1 \times \text{etc}…
\]

\[
p(a, f | e) = z^{-1} \times g^{\lambda_1} \times w^{\lambda_2} \times t_1^{\lambda_3} \times \text{etc}…
\]

\[
p(a, f | e) = \frac{\exp \sum_m \lambda_m h_m(f, a, e; \theta_m)}{\exp(Z)}
\]
Semi-Supervised Training

• Define a semi-supervised algorithm which alternates increasing likelihood with decreasing error
  – Increasing likelihood is similar to EM
  – Discriminatively bias EM to converge to a local maxima of likelihood which corresponds to “better” alignments
    • “Better” = higher $F_\alpha$-score on small gold standard corpus
The EMD Algorithm

Bootstrap

Initial sub-model parameters

D-Step

Tuned lambda vector

Sub-model parameters

E-Step

Viterbi alignments

Translation

M-Step

Viterbi alignments

Tuned lambda vector

Initial sub-model parameters

M-Step

Sub-model parameters

E-Step

Viterbi alignments

Translation

The EMD Algorithm

Alex Fraser
Discussion

• Usual formulation of semi-supervised learning: “using unlabeled data to help supervised learning”
  – Build initial supervised system using labeled data, predict on unlabeled data, then iterate
  – But we do not have enough gold standard word alignments to estimate parameters directly!

• EMD allows us to train a small number of important parameters discriminatively, the rest using likelihood maximization, and allows interaction
  – Similar in spirit (but not details) to semi-supervised clustering
Experiments

• French/English
  – LDC Hansard (67 M English words)
  – MT: Alignment Templates, phrase-based

• Arabic/English
  – NIST 2006 task (168 M English words)
  – MT: Hiero, hierarchical phrases
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>French/English</th>
<th>Arabic/English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-Measure (α = 0.4)</td>
<td>BLEU (1 ref)</td>
</tr>
<tr>
<td>IBM Model 4 (GIZA++) and heuristics</td>
<td>73.5</td>
<td>30.63</td>
</tr>
<tr>
<td>EMD (ACL 2006 model) and heuristics</td>
<td>74.1</td>
<td>31.40</td>
</tr>
<tr>
<td>LEAF+EMD</td>
<td>76.3</td>
<td>31.86</td>
</tr>
</tbody>
</table>
Contributions

- Found a metric for measuring alignment quality which correlates with MT quality
- Designed LEAF, the first generative model of M-to-N discontiguous alignments
- Developed a semi-supervised training algorithm, the EMD algorithm
- Obtained large gains of 1.2 BLEU and 2.8 BLEU points for French/English and Arabic/English tasks
Thank You!