Ensemble Decoding

Domain Adaptation

Multi-metric optimization

Pivot language triangulation
Ensemble Decoding
But first, Mixtures

Translation models: \( m = 1 \ldots M \)

Log-linear mixture:

\[
p(\bar{e}|\bar{f}) \propto \exp \left( \sum_{m}^{M} \lambda_m \log p_m(\bar{e}|\bar{f}) \right)
\]

- Each (phrase-table) component in the usual discriminative SMT model is a mixture.
- The mixture weights are tuned on a dev set.
Linear Mixtures

\[ p(\bar{e} | \bar{f}) = \sum_{m}^{M} \lambda_m p_m(\bar{e} | \bar{f}) \]

- Extract joint phrase pair distribution \( p(e, f) \)
- Find the weights that minimize the cross-entropy of the mixture \( p(e | f) \) with respect to \( p(e, f) \)

\[ \hat{\lambda} = \text{argmax}_\lambda \sum_{\bar{e}, \bar{f}} \tilde{p}(\bar{e}, \bar{f}) \log \sum_{m}^{M} \lambda_m p_m(\bar{e} | \bar{f}) \]
Linear Mixtures

\[ \hat{\lambda} = \arg \max_{\lambda} \sum_{\bar{e}, \bar{f}} \hat{p}(\bar{e}, \bar{f}) \log \sum_{m} \lambda_m p_m(\bar{e} | \bar{f}) \]

- Train the weights on the dev set using any optimization technique (L-BFGS).
- Linear mixtures are used as feature functions in standard discriminative SMT.
- State of the art for domain adaptation in SMT (Foster et al, EMNLP 2010).
Ensemble Decoding

- Previous mixtures of translation models were pre-processing steps.
- This work: Explore mixtures of translation models in the decoder.
- On the fly combination of models in Hiero
1.5 Beihan / North Korea

3.0 bangjiao / diplomatic relations

0.5 yu X₁ you X₂ / have X₂ with X₁

1.5 yu X₁ you X₂ / with X₁ have X₂

\( (\mathbb{R} \cup \{\infty\}, \min, +, \infty, 0) \)

1. have dipl. relns. with N. K = 4.0
2. with N.K. have dipl. relns. = 6.0
• min(4.0, 6.0) option #1 wins
P(have dipl. relns. with N.K | yu Beihan you bangjiao)

\[ \propto \exp \left( \sum_i w_i \phi_i (\tilde{e}, \tilde{f}) \right) \]

\[ w \cdot \phi \]
Ensemble Decoding

\[
P(\text{have dipl. relns. with N.K} \mid \text{yu Beihan you bangjiao})
\]

\[
\propto \exp \left( \frac{w_1 \cdot \phi_1}{1^{st} \text{ model}} \times \frac{w_2 \cdot \phi_2}{2^{nd} \text{ model}} \times \cdots \right)
\]
Ensemble Operations
Weighted Sum (wsum)

\[ p(\bar{e} \mid \bar{f}) \propto \sum_{m}^{M} \lambda_m \exp \left( w_m \cdot \phi_m \right) \]

- Ensemble score is the weighted sum of individual model scores
- \( m \) is each component model, a total of \( M \) components in the ensemble.

<table>
<thead>
<tr>
<th></th>
<th>ml</th>
<th>m2</th>
<th>ens</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>16.5</td>
<td>3.5</td>
<td>10</td>
</tr>
<tr>
<td>fr</td>
<td>4.5</td>
<td>10.5</td>
<td>7.25</td>
</tr>
</tbody>
</table>
Weighted Max (wmax)

\[ p(\bar{e} | \bar{f}) \propto \max_m (\lambda_m \exp (w_m \cdot \phi_m)) \]

- Ensemble score is the weighted max of all the model scores
- The n-best list can contain entries from different models

<table>
<thead>
<tr>
<th></th>
<th>fr</th>
<th>en</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1</td>
<td>4.5</td>
<td>or other disease hereditary metabolic</td>
</tr>
<tr>
<td>m2</td>
<td>10.5</td>
<td>or another hereditary metabolic disease</td>
</tr>
<tr>
<td>ens</td>
<td>4.5</td>
<td>une autre maladie métabolique héréditaire</td>
</tr>
<tr>
<td>ens</td>
<td>3.5</td>
<td></td>
</tr>
</tbody>
</table>
Model Switching (Switch)

\[ p(\bar{e} \mid \bar{f}) = \sum_{m}^{M} \delta(\bar{f}, m) \ p_m(\bar{e} \mid \bar{f}) \]

- Switch models in each CKY cell. Possibly picking a different model from the ensemble.
- The n-best list can contain entries from only one model.

\[ \delta(\bar{f}, m) = \begin{cases} 
1, & m = \text{argmax}_{n \in M} \psi(\bar{f}, n) \\
0, & \text{otherwise}
\end{cases} \]
Model Switching (Switch)

\[
\delta(\vec{f}, m) = \begin{cases} 
1, & m = \arg\max_{n \in M} \psi(\vec{f}, n) \\
0, & \text{otherwise}
\end{cases}
\]

For each cell the model that has the highest weighted score wins:

\[
\psi(\vec{f}, n) = \lambda_n \max_{\vec{e}} (w_n \cdot \phi_n(\vec{e}, \vec{f}))
\]

For each cell, the model with highest weighted sum of scores wins:

\[
\psi(\vec{f}, n) = \lambda_n \exp \left( w_n \cdot \phi_n(\vec{e}, \vec{f}) \right)
\]
Product (prod)

\[ p(\bar{e} \mid \bar{f}) \propto \exp \left( \sum_{m}^{M} \lambda_m (w_m \cdot \phi_m) \right) \]

- Compute the product of all the probabilities in the ensemble (sum of log-probs).
- A Logarithmic Opinion Pool (LOP).
- LOPs work best when the ensemble is used to down-vote a highly confident but incorrect candidate.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>m1</td>
<td>1.5</td>
</tr>
<tr>
<td>m2</td>
<td>10.5</td>
</tr>
<tr>
<td>m3</td>
<td>12.0</td>
</tr>
<tr>
<td>ens</td>
<td>24.0</td>
</tr>
</tbody>
</table>

fr | une autre maladie métabolique héréditaire
en | or other disease hereditary metabolic
Domain Adaptation
Domain Adaptation

<table>
<thead>
<tr>
<th>Domain</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMEA (Medical)</td>
<td>11770</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OUT-of-domain</td>
<td></td>
<td></td>
<td>1522</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Train</th>
</tr>
</thead>
<tbody>
<tr>
<td>EuroParl (fr-en)</td>
<td>1.3M</td>
</tr>
</tbody>
</table>
Domain Adaptation

• Scaling the model scores using ensemble weights:
  • Find the appropriate model scores that can participate in an ensemble.
  • We use CONDOR (Vanden Berghen and Bersini, 2005) which uses Powell’s algorithm and no gradient information.
  • Component weights for each mixture operation is tuned on the dev set.
Domain Adaptation Results

- **WMAX**: Uniform 35.39, Tuned 35.47, Normalized 35.47
- **WSUM**: Uniform 35.35, Tuned 35.53, Normalized 35.45
- **SWITCH:MAX**: Uniform 35.93, Tuned 35.96, Normalized 35.96
- **SWITCH:SUM**: Uniform 34.9, Tuned 34.9, Normalized 34.9
- **PROD**: Uniform 33.93, Tuned 35.02, Normalized 35.24

**IN+OUT: 33.76**

**LINMIX: 35.57**

Significant at s=0.01
Example

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>aménorrhée , menstruations irrégulières</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF</td>
<td>amenorrhoea , irregular menstruation</td>
</tr>
<tr>
<td>IN</td>
<td>amenorrhoea , menstruations irrégulières</td>
</tr>
<tr>
<td>OUT</td>
<td>aménorrhée , irregular menstruation</td>
</tr>
<tr>
<td>ENSEMBLE</td>
<td>amenorrhoea , irregular menstruation</td>
</tr>
</tbody>
</table>
Example

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>le traitement par naglazyme doit être supervisé par un médecin ayant l'expérience de la prise en charge des patients atteints de mps vi ou d’une autre maladie métabolique héréditaire.</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF</td>
<td>naglazyme treatment should be supervised by a physician experienced in the management of patients with mps vi or other inherited metabolic diseases.</td>
</tr>
<tr>
<td>IN</td>
<td>naglazyme treatment should be supervisée by a doctor the with in the management of patients with mps vi or other hereditary metabolic disease.</td>
</tr>
<tr>
<td>OUT</td>
<td>naglazyme’s treatment must be supervised by a doctor with the experience of the care of patients with mps vi. or another disease hereditary metabolic.</td>
</tr>
<tr>
<td>ENSEMBLE</td>
<td>naglazyme treatment should be supervised by a physician experienced in the management of patients with mps vi or other hereditary metabolic disease.</td>
</tr>
</tbody>
</table>
Multi-metric optimization

Joint work with Baskaran Sankaran and Kevin Duh
Multi-metric optimization

• Quite a few proposals for MT evaluation.
• In this talk, the focus is on BLEU, RIBES, TER, METEOR.
• Can other metrics be useful as a loss function for training SMT systems. (be useful how?)
• Most systems tune towards BLEU and test on BLEU. Can other metrics provide a second opinion?
Multi-objective optimization

\[ \max_w (F_1(w), F_2(w), \ldots, F_k(w)) \]

• Find one \( w \) that simultaneously optimizes \( k \) objectives.

• A well formed notion of optimality wrt multiple objectives: Pareto optimality
Finding Pareto points

• Duh et al (ACL 2012) give an algorithm called PMO-PRO that finds Pareto optimal points as part of the tuning step.

• PRO (May and Hopkins, EMNLP 2011) show a pairwise ranking classifier can be used to train an SMT log-linear model.

• PMO-PRO puts Pareto points as positive examples and low scoring non-Pareto points as negative examples.

• This can be used to find Pareto points in the dev set.
Using the Pareto Points

• Each Pareto point in the dev set is a weight vector that produced that point.

• **PMO-Ensemble**: Each of these weight vectors is a model and we can simply combine them using an ensemble model.

• **Union**: Take the union of “good” points wrt multiple objectives as positive examples and vice versa for negative examples. Simpler version of PMO-PRO.
Using the Pareto Points

\[
\max_w (F_1(w), F_2(w), \ldots, F_k(w))
\]

- **Ensemble Tuning:**
  - For each \(F_i(w)\) perform error rate tuning (using PRO) to obtain the best \(w_i\) according to \(F_i\) in each iteration of tuning.
  - When decoding the dev set for the next search for \(w\) use an ensemble model with the same features but weights: \(w_1, ..., w_k\)
  - Tune the ensemble model hyperparameters using PMO-PRO to get Pareto points in the ensemble.
  - Repeat.
Can multi-metric tuning help a single metric?

Single Objective

Ensemble Tuning: 2 Metrics

Ensemble Tuning: 3 Metrics

>3 Metrics
Pivot language triangulation
Triangulation

- **Direct**: From source to target using available data
- Phrase-based triangulation (Cohn and Lapata, ACL 2007)
Phrase-based Triangulation

- Create a triangulated phrase table using source, pivot and target language data.

\[ p(e|f) = \sum_i p(e, i|f) \]
\[ = \sum_i p(e|i, f)p(i|f) \]
\[ \approx \sum_i p(e|i)p(i|f) \]

- **Mixture**: Interpolate the triangulated model with the direct source to target model
Ensemble-based Triangulation

• Typically, one pivot language does not provide an improvement.

• More pivot languages used, the better.

• Ensemble-based Triangulation: an ensemble of different pivot models.

• Each one goes from source to pivot\textsubscript{k} to target for pivot\textsubscript{1} to pivot\textsubscript{M}

• **Ensemble**: Finally add the direct source to target model to the ensemble as well.
Experiment

• Compare Direct, Mixture and Ensemble

• Use EuroParl (en, fr, de, es, it)

• Each source language is translated to a target language through 3 pivot languages.

• For example, en to fr goes through de, es, it

• 10K sentence pairs (as in Cohn and Lapata, ACL 2007)

• To be done: 700K EuroParl corpus.
### Ensemble Triangulation

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>de</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>-</td>
<td>+0.2</td>
<td>+1.0</td>
<td>-0.09</td>
</tr>
<tr>
<td>es</td>
<td>+0.24</td>
<td>-</td>
<td>+1.06</td>
<td>+0.38</td>
</tr>
<tr>
<td>fr</td>
<td>+0.9</td>
<td>+0.93</td>
<td>-</td>
<td>+0.03</td>
</tr>
<tr>
<td>de</td>
<td>+0.75</td>
<td>-0.48</td>
<td>+0.06</td>
<td>-</td>
</tr>
</tbody>
</table>

Comparison of Ensemble Model and Mixture model (Cohn and Lapata, ACL 2007)
Summary

• Ensemble models combine translation models during SMT decoding.
  • Allows more dynamic combination methods.
  • Do not need to be tuned (with uniform weights)

• Applied to:
  • Domain adaptation
  • Multi-metric optimization
  • Pivot language triangulation
<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Direct</th>
<th>Mixture</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-de</td>
<td>17.66</td>
<td>18.09</td>
<td>18</td>
</tr>
<tr>
<td>en-es</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en-fr</td>
<td>24.83</td>
<td>24.39</td>
<td>25.39</td>
</tr>
<tr>
<td>es-en</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>es-fr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>es-de</td>
<td>16.43</td>
<td>17.77</td>
<td>18.15</td>
</tr>
<tr>
<td>fr-de</td>
<td>16.84</td>
<td>17.68</td>
<td>17.71</td>
</tr>
<tr>
<td>fr-en</td>
<td>28.97</td>
<td>29.15</td>
<td>30.05</td>
</tr>
<tr>
<td>fr-es</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de-en</td>
<td>22.03</td>
<td>22.78</td>
<td>22.78</td>
</tr>
<tr>
<td>de-es</td>
<td>20.77</td>
<td>22.57</td>
<td>22.09</td>
</tr>
<tr>
<td>de-fr</td>
<td>17.47</td>
<td>18.53</td>
<td>18.59</td>
</tr>
</tbody>
</table>