MT System Combination

11-731
Machine Translation
Alon Lavie
April 9, 2015

With acknowledged contributions from Silja Hildebrand and Kenneth Heafield
Goals and Challenges

- Different MT systems have different strengths and weaknesses
  - Different approaches: Phrase-based, Hierarchical, Syntax-based, RBMT, EBMT
  - Different domains, training data, tuning data

- **Scientific Challenge:**
  - How to combine the output of multiple MT engines into a selected output that outperforms the originals in translation quality?

- **Selecting the best output** on a sentence-by-sentence basis (classification), or a more synthetic combination?

- Range of approaches to address the problem
- Can result in very significant gains in performance
Several Different MT System Outputs

Reference Translation:
*hoffman was addicted to drugs, fortunately awaking in a timely manner to begin an acting career*

- hoffman was obsessed timely wake up to create a career drug
- hoffman were drug fortunately awakenin in a timely manner to create career
- hoffman previously enamored drug, luckily i realized create career
- hoffman was mesmerized by drug but woke up in a timely manner to create career
- hoffmann was obsessed drug, in a timely manner to create a career
- hoffman has fortunately drug come to realize in a timely manner for performing arts to open up the cause

*Chinese-English MT06*

- Statistical Phrase Based  ➔  Statistical Hierarchical  ➔  Example Based

Translation hypotheses are in order of the systems testset BLEU score
Combination Architecture

- Parallel Combination
  - Run multiple MT systems in parallel, then select or combine their outputs
- Serial Combination
  - Second stage decoding using a different approach
- Model Combination
  - Train separate models, then combine them for joint decoding
Parallel Combination
Serial Combination

Source Language Text → MT System → MT System → Translation
Model Combination

Source Language Text

Decoder

Phrase Table Combination

Lexicon Combination

Reordering Combination

Phrase Table S1

Phrase Table S2

Lexicon S1

Lexicon S2

Reordering S1

Reordering S2

Target Language Text
Main Approaches

- **Parallel Combination:**
  - Hypothesis Selection approaches
  - Lattice Combination
  - Confusion (or Consensus) Networks
  - Alignment-based Synthetic Multi-Engine MT (MEMT)

- **Serial Combination:**
  - Automated Post-Editing: RBMT + SMT
  - Cross combinations of parallel combinations (GALE)

- **Model Combination:**
  - Combine lexica, phrase tables, LMs
  - Ensemble decoding (Sarkar et al, 2012)
Hypothesis Selection Approaches

- **Main Idea**: construct a classifier that given several translations for the same input sentence selects the “best” translation (on a sentence-by-sentence basis)
- Should “beat” a baseline of always picking the system that is best in the aggregate
- Main knowledge sources for scoring the individual translations are standard statistical target-language LMs, confidence scores for each engine, consensus information
- Examples:
  - [Tidhar & Kuessner, 2000]
  - [Hildebrand and Vogel, 2008]
Hypothesis Selection

Source Language Text → MT System → MT System Combination → Translation
Hypothesis Selection

- Work here at CMU (InterACT) by Silja Hildebrand:
  - Combines n-best lists from multiple MT systems and re-ranks them with a collection of computed features
  - Log-linear feature combination is independently tuned on a development set for max-BLEU
  - Richer set of features than previous approaches, including:
    - Standard n-gram LMs (normalized by length)
    - Lexical Probabilities (from GIZA statistical lexicons)
    - Position-dependent n-best list word agreement
    - Position-independent n-best list n-gram agreement
    - N-best list n-gram probability
    - Aggregate system confidence (based on BLEU)
  - Applied successfully in GALE and WMT-09
  - Improvements of 1-2 BLEU points above the best individual system on average
  - Complimentary to other approaches – is used to select “back-bone” translation for confusion network in GALE
Position-Dependent Word Agreement

n-best list for one source sentence

- $t = 0$ agreement: 30%
- $t = 1$ agreement: 50%
- $t = 2$ agreement: 70%
Position-Independent Word Agreement

n-best list for one source sentence

- $e_5$

- $n = 1$
  - word agreement: 90%

- $n = 3$
  - tri-gram agreement: 50%

- $n = 5$
  - 5-gram agreement: 30%

Agreement score for $n = 1$ to $6$ as separate features
N-gram Agreement vs. N-gram Probability

San Francisco

n = 2
bi-gram agreement: 0.6%

P( Francisco | San ) = 3 / 3
bi-gram probability: 100%

LM n-gram probability gives information on word order.
Lattice-based MEMT

- Earliest approach, first tried in CMU’s PANGLOSS in 1994, and still active in recent work

- Main Ideas:
  - Multiple MT engines each produce a lattice of scored translation fragments, indexed based on source language input
  - Lattices from all engines are combined into a global comprehensive lattice
  - Joint Decoder finds best translation (or n-best list) from the entries in the lattice
# Lattice-based MEMT: Example

<table>
<thead>
<tr>
<th>El punto de descargo</th>
<th>se cumplirá en</th>
<th>el puente Agua Fria</th>
</tr>
</thead>
<tbody>
<tr>
<td>The drop-off point</td>
<td>will comply with</td>
<td>The cold Bridgewater</td>
</tr>
<tr>
<td>El punto de descargo</td>
<td>se cumplirá en</td>
<td>el puente Agua Fria</td>
</tr>
<tr>
<td>The discharge point</td>
<td>will self comply in</td>
<td>the “Agua Fria” bridge</td>
</tr>
<tr>
<td>El punto de descargo</td>
<td>se cumplirá en</td>
<td>el puente Agua Fria</td>
</tr>
<tr>
<td>Unload of the point</td>
<td>will take place at</td>
<td>the cold water of bridge</td>
</tr>
</tbody>
</table>
Lattice-based MEMT

- Main Drawbacks:
  - Requires MT engines to provide lattice output → often difficult to obtain!
  - Lattice output from all engines must be compatible: common indexing based on source word positions → difficult to standardize!
  - Common TM used for scoring edges may not work well for all engines
  - Decoding does not take into account any reinforcements from multiple engines proposing the same translation for any portion of the input
Consensus Network Approach

• **Main Ideas:**
  - Collapse the collection of linear strings of multiple translations into a minimal consensus network (“sausage” graph) that represents a finite-state automaton
  - Edges that are supported by multiple engines receive a score that is the sum of their contributing confidence scores
  - Decode: find the path through the consensus network that has optimal score
  - Examples:
    - [Bangalore et al, 2001]
    - [Rosti et al, 2007]
Fig. 4. Lattice representation of the result of the multiple alignment. The weights on the arcs are negative logarithm of the probability that word.
Confusion Network Approaches

• Similar in principle to the Consensus Network approach
  – Collapse the collection of linear strings of multiple translations into minimal confusion network(s)
• Main Ideas and Issues:
  – Aligning the words across the various translations:
    • Can be aligned using TER, ITGs, statistical word alignment
  – Word Ordering: picking a “back-bone” translation
    • One backbone? Try each original translation as a backbone?
  – Decoding Features:
    • Standard n-gram LMs, system confidence scores, agreement
  – Decode: find the path through the consensus network that has optimal score
• Developed and used extensively in GALE (also WMT)
• Nice gains in translation quality: 1-4 BLEU points
Confusion Network Construction

Align Words, Build Confusion Network

- hoffman was obsessed timely wake up to create a career drug
- hoffman were drug fortunately awakening in a timely manner to create career
- hoffman previously enamored drug, luckily i realized create career
- hoffman was mesmerized by drug but woke up in a timely manner to create career
- hoffmann was obsessed drug in a timely manner to create a career
- hoffman has fortunately drug come to realize in a timely manner for performing arts to open up the cause
Confusion Network Decoding

I like eating chocolate ice-cream.
I like to eat vanilla ice-cream.
I like to eat ice-cream with chocolate.
I like ice-cream.

I like ε eating chocolate ε ice-cream.
I like to eat vanilla ε ice cream.

choose as skeleton

skeleton determines word order

I like to eat chocolate ice-cream.
Confusion Networks - Challenges

- Word alignment
  - TER alignment (Translation Edit Rate)
  - ITG based alignment (Inversion Transduction Grammar) - invWER
  - Use morphology, synonyms, POS tag
  - Go to phrases
    - Difficult without source-target phrase alignment available
- Double translations
- Dropped words
- Pairwise vs. incremental alignment
  - Next hypothesis is aligned to the existing network, not to the skeleton
  - Order of adding hypothesis does make a difference, e.g. use increasing TER/decreasing BLEU of the system
CMU’s Alignment-based Multi-Engine System Combination

- Works with any MT engines
  - Assumes original MT systems are “black-boxes” – no internal information other than the translations themselves
- Explores broader search spaces than other MT system combination approaches using linguistically-based and statistical features
- Achieves state-of-the-art performance in research evaluations over past couple of years
- Developed over last ten years under research funding from several government grants (DARPA, DoD and NSF)
Alignment-based MEMT

Two Stage Approach:

1. Identify common words and phrases across the translations provided by the engines
2. Decode: search the space of synthetic combinations of words/phrases and select the highest scoring combined translation

Example:

1. announced afghan authorities on saturday reconstituted four intergovernmental committees
2. The Afghan authorities on Saturday the formation of the four committees of government
Alignment-based MEMT

Two Stage Approach:
1. Identify common words and phrases across the translations provided by the engines
2. Decode: search the space of synthetic combinations of words/phrases and select the highest scoring combined translation

Example:
1. announced *afghan authorities on saturday* reconstituted *four intergovernmental committees*
2. The *Afghan authorities on Saturday* the formation of the *four committees of government*

MEMT: the *afghan authorities announced on Saturday* the formation of *four intergovernmental committees*
The String Alignment Matcher

- Developed as a component in the METEOR Automatic MT Evaluation metric
- Finds maximal alignment match with minimal “crossing branches”
- Allows alignment of:
  - Identical words
  - Morphological variants of words
  - Synonymous words (based on WordNet synsets)
  - Paraphrases
- Implementation: approximate single-pass search algorithm for best match using pruning of sub-optimal sub-solutions
MEMT Alignment

Match surface, stems, WordNet synsets, and automatic paraphrases.
Minimize crossing alignments.

Twice that produced by nuclear plants.
Double that that produce nuclear power stations.

The MEMT Decoder Algorithm

- Algorithm builds collections of partial hypotheses of increasing length
- Partial hypotheses are extended by selecting the “next available” word from one of the original systems
- Sentences are assumed mostly synchronous:
  - Each word is either *aligned* with another word or is an *alternative* of another word
- Extending a partial hypothesis with a word “pulls” and “uses” its aligned words with it, and marks its alternatives as “used”
- Partial hypotheses are scored and ranked
- Pruning and re-combination
- Hypothesis can end if any original system proposes an end of sentence as next word
Decoding Example

System 1: Now can know why.
System 2: Now we can now know why.

Partial Hypothesis

\[
\begin{align*}
\text{Now} \\
\text{Now} \\
\end{align*}
\]
Decoding Example

System 1: Now can know why.

System 2: Now we can now know why.

Partial Hypothesis
Decoding Example

System 1: Now can know why.

System 2: Now we can now know why.

Partial Hypothesis:

\[
\begin{cases}
\text{can} \\
\text{can}
\end{cases}
\]
Decoding Example

System 1: Now can know why.

System 2: Now we can now know why.

Partial Hypothesis

\{ know, now \}
Scoring MEMT Hypotheses

- **Features:**
  - N-gram Language Model score based on filtered large-scale target language LM
  - OOV feature
  - N-gram support features with n-grams matches from the original systems (unigrams to 4-grams)
  - Length

- **Scoring:**
  - Weighed Log-linear feature combination tuned on development set
  - Weights are tuned using MERT on a held-out tuning set
N-gram Match Support Features

**System 1:** Supported Proposal of France

**System 2:** Support for the Proposal of France

**Hypothesis:** Support for Proposal of France

<table>
<thead>
<tr>
<th></th>
<th>Unigram</th>
<th>Bigram</th>
<th>Trigram</th>
<th>Quadgram</th>
</tr>
</thead>
<tbody>
<tr>
<td>System 1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>System 2</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Hyper-Parameters

• Selecting among the various MT systems available for combination
  – Combine all or just a subset?
  – Criteria for selection: metric scores, diversity of approach, other...

• Internal Hyper-settings:
  – “Horizon”: when to drop lingering words
  – N-gram match support features: per individual system or aggregate across systems?

• Highly efficient implementation allows executing exhaustive collection of experiments with different hyper-parameter settings on distributed parallel high-computing clusters
Performance Results on NIST-2009 and WMT-2009

<table>
<thead>
<tr>
<th>Source</th>
<th>Top</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>58.55</td>
<td>+6.67</td>
</tr>
<tr>
<td>Czech</td>
<td>21.98</td>
<td>+0.80</td>
</tr>
<tr>
<td>French</td>
<td>31.56</td>
<td>+0.42</td>
</tr>
<tr>
<td>German</td>
<td>23.88</td>
<td>+2.57</td>
</tr>
<tr>
<td>Hungarian</td>
<td>13.84</td>
<td>+1.09</td>
</tr>
<tr>
<td>Spanish</td>
<td>28.79</td>
<td>+0.10</td>
</tr>
<tr>
<td>Urdu</td>
<td>34.72</td>
<td>+1.84</td>
</tr>
</tbody>
</table>

Table: Post-evaluation uncased BLEU gains on NIST and WMT tasks.
Performance Results on WMT-2010

<table>
<thead>
<tr>
<th>French-English</th>
<th>589–716 judgments per combo</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>≥others</td>
</tr>
<tr>
<td>RWTH-COMBO</td>
<td>0.77</td>
</tr>
<tr>
<td>CMU-HYP-COMBO</td>
<td>0.77</td>
</tr>
<tr>
<td>DCU-COMBO</td>
<td>0.72</td>
</tr>
<tr>
<td>LIUM</td>
<td>0.71</td>
</tr>
<tr>
<td>CMU-HEA-COMBO</td>
<td>0.70</td>
</tr>
<tr>
<td>UPV-COMBO</td>
<td>0.68</td>
</tr>
<tr>
<td>NRC</td>
<td>0.66</td>
</tr>
<tr>
<td>CAMBRIDGE</td>
<td>0.66</td>
</tr>
<tr>
<td>UEDIN</td>
<td>0.65</td>
</tr>
<tr>
<td>LIMSI</td>
<td>0.65</td>
</tr>
<tr>
<td>JHU-COMBO</td>
<td>0.65</td>
</tr>
<tr>
<td>RALI</td>
<td>0.65</td>
</tr>
<tr>
<td>LIUM-COMBO</td>
<td>0.64</td>
</tr>
<tr>
<td>BBN-COMBO</td>
<td>0.64</td>
</tr>
<tr>
<td>RWTH</td>
<td>0.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English-French</th>
<th>740–829 judgments per combo</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>≥others</td>
</tr>
<tr>
<td>RWTH-COMBO</td>
<td>0.75</td>
</tr>
<tr>
<td>CMU-HEA-COMBO</td>
<td>0.74</td>
</tr>
<tr>
<td>UEDIN</td>
<td>0.70</td>
</tr>
<tr>
<td>KOC-COMBO</td>
<td>0.68</td>
</tr>
<tr>
<td>UPV-COMBO</td>
<td>0.66</td>
</tr>
<tr>
<td>RALI</td>
<td>0.66</td>
</tr>
<tr>
<td>LIMSI</td>
<td>0.66</td>
</tr>
<tr>
<td>RWTH</td>
<td>0.63</td>
</tr>
<tr>
<td>CAMBRIDGE</td>
<td>0.63</td>
</tr>
</tbody>
</table>
Performance Results on WMT-2010

### Spanish-English

<table>
<thead>
<tr>
<th>System</th>
<th>≥others</th>
</tr>
</thead>
<tbody>
<tr>
<td>UEDIN</td>
<td>0.69</td>
</tr>
<tr>
<td>CMU-HEA-COMBO</td>
<td>0.66</td>
</tr>
<tr>
<td>UPV-COMBO</td>
<td>0.66</td>
</tr>
<tr>
<td>BBN-COMBO</td>
<td>0.62</td>
</tr>
<tr>
<td>JHU-COMBO</td>
<td>0.55</td>
</tr>
<tr>
<td>UPC</td>
<td>0.51</td>
</tr>
</tbody>
</table>

(1385–1535 judgments per combo)

### English-Spanish

<table>
<thead>
<tr>
<th>System</th>
<th>≥others</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU-HEA-COMBO</td>
<td>0.68</td>
</tr>
<tr>
<td>KOC-COMBO</td>
<td>0.62</td>
</tr>
<tr>
<td>UEDIN</td>
<td>0.61</td>
</tr>
<tr>
<td>UPV-COMBO</td>
<td>0.60</td>
</tr>
<tr>
<td>RWTH-COMBO</td>
<td>0.59</td>
</tr>
<tr>
<td>DFKI</td>
<td>0.55</td>
</tr>
<tr>
<td>JHU</td>
<td>0.55</td>
</tr>
<tr>
<td>UPV</td>
<td>0.55</td>
</tr>
<tr>
<td>CAMBRIDGE</td>
<td>0.54</td>
</tr>
<tr>
<td>UPV-NNLM</td>
<td>0.54</td>
</tr>
</tbody>
</table>

(516–673 judgments per combo)
### Human Evaluation Results

<table>
<thead>
<tr>
<th>Language Combination</th>
<th>BLEU</th>
<th>Top Individual BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Czech-English</strong></td>
<td>0.64</td>
<td>0.62</td>
</tr>
<tr>
<td><strong>German-English</strong></td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td><strong>Spanish-English</strong></td>
<td>0.71</td>
<td>0.67</td>
</tr>
<tr>
<td><strong>French-English</strong></td>
<td>0.67</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Clean Haitian-English</strong></td>
<td>0.52</td>
<td>0.48</td>
</tr>
<tr>
<td><strong>English-Czech</strong></td>
<td>0.53</td>
<td>0.61</td>
</tr>
<tr>
<td><strong>English-German</strong></td>
<td>0.61</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>English-Spanish</strong></td>
<td>0.69</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>English-French</strong></td>
<td>0.69</td>
<td>0.51</td>
</tr>
<tr>
<td><strong>Raw Haitian-English</strong></td>
<td>0.69</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Smoothing MERT in SMT  
[Cettolo, Bertoldi and Federico 2011]

- Interesting application of MT system combination to overcome instability of MERT optimization in SMT
  - Perform MERT multiple times
  - Use the CMU MEMT system to combine the different instances of the same MT system

<table>
<thead>
<tr>
<th>en–fi</th>
<th>BLEU%</th>
<th>stdev</th>
<th>[min,max]</th>
</tr>
</thead>
<tbody>
<tr>
<td>optSample</td>
<td>35.95</td>
<td>0.080</td>
<td>[35.83,36.07]</td>
</tr>
<tr>
<td>avg6</td>
<td>35.97</td>
<td>0.023</td>
<td>[35.93,36.01]</td>
</tr>
<tr>
<td>sysComb6</td>
<td>36.34</td>
<td>0.106</td>
<td>[36.21,36.50]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>el–fr</th>
<th>BLEU%</th>
<th>stdev</th>
<th>[min,max]</th>
</tr>
</thead>
<tbody>
<tr>
<td>optSample</td>
<td>58.22</td>
<td>0.104</td>
<td>[58.01,58.33]</td>
</tr>
<tr>
<td>avg6</td>
<td>58.09</td>
<td>0.043</td>
<td>[58.02,58.15]</td>
</tr>
<tr>
<td>sysComb6</td>
<td>58.92</td>
<td>0.114</td>
<td>[58.71,59.08]</td>
</tr>
</tbody>
</table>

Table 4: Results for the ACQUIS task on the test set.
CMU MEMT System is Open Source

- http://kheafield.com/code/memt/
- Open Source, LGPL license
- Freely available for research and commercial use
References

Questions?