SYNTAX FOR

STATISTICAL MACHINE TRANSLATION

Franz Josef Och

och@isi.edu

Franz Josef Och – USC/ISI
Daniel Gildea – UPenn
Anoop Sarkar – SFU
Kenji Yamada – XRCE
Sanjeev Khudanpur – JHU

Alex Fraser – USC/ISI
Shankar Kumar – JHU
David Smith – JHU
Libin Shen – UPenn

Viren Jain – UPenn
Katherine Eng – Stanford
Zhen Jin – Mt. Holyoke

+ Assoc.: Dragomir Radev – Univ. of Michigan
major tasks:

- **MODELING**: introducing structures into the probabilistic dependencies

\[
( \frac{I}{I} | \frac{J}{J} ) = \lambda^M_{1}( \frac{I}{I} | \frac{J}{J} )
\]

- **TRAINING**: estimation of free parameters using training data \( (f^S_1, e^S_1) \)

\[
\lambda^M_{1} = \arg\max_{\lambda^M_{1}} \text{SOME-CRITERION}(f^S_1, e^S_1, M)
\]

- **SEARCH**: finding 'best' translation

\[
\hat{e} = \arg\max_{e} \hat{\lambda}^M_{1}(e|f)
\]

additional important tasks:

- data acquisition: collection of parallel training data
- pre-/postprocessing: tokenization, normalization, ...
- evaluation: assessing the quality of MT output
Statistical Machine Translation: Log–linear models

Approach to model \((e|f)\): decompose into simpler dependencies

- \(m(I_1 J_1)\) with \(m = 1\) : feature functions
- \(m\) with \(m = 1\) : interpolation parameter

Log–linear models:

\[
\begin{align*}
\lambda^M_{I_1}(I_1 J_1) & = \prod_{m=1}^{M} m(I_1 J_1) \\
& = \frac{\exp[\sum_{m=1}^{M} m(I_1 J_1) \cdot m]}{\sum_{e I_1} \exp[\sum_{m=1}^{M} m(I_1 J_1) \cdot m]}
\end{align*}
\]

Very simple decision rule (with zero-one loss function):

\[
\hat{I}_1^I = \arg\max_{e I_1} \left\{ \sum_{m=1}^{M} m(I_1 J_1) \cdot m \right\}
\]

What are the relevant dependencies?
basic idea: remember ALL seen “bilingual phrase” together with word alignment

→ bilingual phrase pair with alignment: alignment template

→ restrict length (here: up to 7 words)

→ use word classes for generalization
hidden variables of alignment template translation model:

- \( K_1 \): sequence of alignment templates
- \( K_1 \): alignment of alignment templates

Feature functions:

alignment template selection:

\[ \text{AT}(I_1 J_1 K_1 K_1) = \log \left( \frac{K_1}{I_1} \right) = \log \prod_{k=1}^{K_1} \left( \frac{k}{\sim k} \right) \]

(phrase) alignment:

\[ \text{AL}(I_1 J_1 K_1 K_1) = \log \left( \frac{K_1}{I_1 J_1 K_1} \right) = \]

word selection:

\[ \text{WRD}(I_1 J_1 K_1 K_1) = \log \prod_{i=1}^{I_1} \left( \frac{i}{\text{aligned French words}} \right) \]

word-based trigram language model feature function:

\[ \text{LM}(I_1 J_1 K_1 K_1) = \log \prod_{i=1}^{I_1} \left( \frac{i}{i-2 \ i-1} \right) \]

number of produced words (word penalty):

number of produced phrases (alignment template penalty):

...
(Discriminative) Training Criteria:

- standard training criterion: maximum posterior probability

\[ \hat{\lambda}_M^1 = \arg\max_{\lambda_1^M} \left\{ \sum_{s=1}^{S} \log \lambda_1^M(e_s | f_s) \right\} \]

- direct optimization of BLEU score / NIST score [Och03]

Search:

- DP beam-search producing sentence from left to right
  baseline feature functions can be easily decomposed for left-to-right search
- restrict possible reordering to a maximal jump distance of 10 words
- N-best list extraction: (optimal) A* search
CHINESE-ENGLISH DARPA/NIST MT EVALUATION (LARGE DATA TRACK)

<table>
<thead>
<tr>
<th>System</th>
<th>2002</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>alignment template approach</td>
<td>7.6 (RWTH)</td>
<td>9.0 (ISI)</td>
</tr>
<tr>
<td>best competing research systems</td>
<td>7.3</td>
<td>7.9</td>
</tr>
<tr>
<td>best of six commercial off-the-shelf systems</td>
<td>6.1</td>
<td>6.6</td>
</tr>
</tbody>
</table>

Note: 95%-confidence intervals: +/- 0.15

Good results, but: system makes often 'stupid' errors
• The resolution urged the to the ceasefire and demands ...

  BETTER: The resolution urges Israel and the Palestinians to cease fire and demands ...

• Indonesia that oppose the presence of foreign troops.

  BETTER: Indonesia reiterated its opposition to foreign military presence.

• in brief ) ( South Korea will be on high-ranking official to visit North Korea

  BETTER: (News in Brief) South Korean high-ranking officials to visit North Korea in April

• Condemns US interference in its internal affairs.

  BETTER: Ukraine condemns US interference in its internal affairs.

• Japan to freeze Rusia to provide humanitarian aid

  BETTER: Japan to freeze humanitarian assistance to Russia
• ... if the west further sanctions against zimbabwe ...
  BETTER: ... if western countries impose further sanctions against zimbabwe ...

• ... he is fully able to activate team .
  BETTER: ... he is fully able to activate the team .

• ... , particularly those who cheat the audience the players.
  BETTER: ... , particularly those players who cheat the audience.

• ... the trial of the outcome ...
  BETTER: ... the outcome of the trial ...
idea: add additional feature functions that 'explicitly' look at syntactic well-formedness of produced translation

syntactic framework: statistical parser for source and target language

feature functions depend also on parse trees: \( (\quad) \rightarrow (e \ f) \)

feature functions: simple \( \rightarrow \) complicated

- verb present in both sentences
- verb takes same number of arguments
- ...
- relationships between phrases should be transferred
- ...
- projecting one tree: constituent reordering
- aligning tree pairs: elementary tree pairs
1. Error Analysis / Feature Hunting
   (a) step 1: detect systematic error by contrasting produced/oracle/reference translations
   (b) step 2: develop feature function for that error
   (c) step 3: perform disc. training with that feature
   (d) step 4: evaluate system performance and keep feature if performance improves
   (e) goto 1

2. Development of Feature Functions
   → ...

3. Diskriminative Training Techniques
   → maximum entropy training: YASMET toolkit
   → directly optimizing BLEU scores: opt-nbest
   → SVM reranking

4. Search:
   → -best rescoring
   → minimum Bayes risk search with syntactically motivated loss functions
What's already done?

- **Framework:** Large Data Track Chinese–English → 150M words (per language)
  
  1. Parallel training data: Chinese treebank, FBIS, Xinhua News, Hongkong News, UN, Hongkong Hansards, Sinorama, Hongkong Laws
  2. Chinese treebank for parser training
  3. Development data: Evalset-01, Devset-03-1 Devset-03-2 → 175K words
  4. Test data: Evalset-02 → 25K words
  5. Blind test data: Evalset-03 → 25K words
  6. 16384-best lists for development and test corpora produced by baseline alignment template system
     
     | -best size | 1 | 1024 | 16384 |
     | BLEU[%]    | 32 | 45%  | 50%   |

  7. Large parts of the data are tagged/chunked/parsed

- **Ready-to-use tools**
  
  1. Syntax tools: English/Chinese POS tagger, chunker, parser + conv. tools
  2. Discriminative training: maximum entropy training (YASMET), maximum BLEU training, maximum NIST training
  3. Feature functions: tree-to-tree/tree-to-string alignment models
How do we define success?

We are successful, if

- we achieve significant improvement of BLEU/NIST scores
  
  absolute improvement of BLEU: 1%: statistically significant
  
  1 − 2%: ok
  
  2 − 3%: good
  
  3 − 4%: very good
  
  4 − 6%: fantastic (more is unrealistic)

  → potential problem: BLEU/NIST are probably not sensitive enough to measure improvements in syntactic wellformedness

- human evaluation judges week-5 system significantly better (with respect to syntax) than initial system

  → plan: subjective evaluation performed at Butler Hill Group or at LDC with five human evaluators for 900 sentences

- ( we show that explicit models of syntax are useless in MT )
• week 1:
  → one-click-program for:
  computing FF, training of parameters on development corpus, BLEU on test corpus
  goal: very quick development training/test cycle of under sixty minutes (for simple FF’s)
  → integrate tree-to-tree, tree-to-string alignment models
  → contrastive error analysis
  → plan development of new feature functions

• week 2/3/4/5/6:
  → error analysis / feature function development
  → end of week: evaluate progress on blind test corpus

• end of week 5:
  → build system for subjective evaluation
Summary:

- Starting point: best existing Chinese-English MT system: alignment template system from USC/ISI

- Contrastive error analysis

- Develop specific syntactic feature functions that try to 'fix' the errors

- Perform discriminative training of feature function weights