WS 2003 Syntax for Statistical Machine Translation
The Team (12 + 1)

- Franz Josef Och - ISI
- Daniel Gildea - Upenn
- Anoop Sarkar - SFU
- Kenji Yamada - XRCE
- Sanjeev Khudanpur - JHU
- + Dragomir Radev - Univ. of Michigan
- Alex Fraser - ISI
- Shankar Kumar - JHU
- David Smith - JHU
- Libin Shen - Upenn
- Viren Jain - Upenn
- Katherine Eng - Stanford
- Zhen Jin - Mt. Holyoke
Statistical Machine Translation

• Enormous progress in MT in recent years due to **statistical approaches**
  – Verbmobil: German-English (speech-to-speech)
  – TIDES project: Chinese, Arabic, Hindi

• Advantages:
  – **Better** quality
  – **Faster** (rapid) development cycle
  – **Cheaper**
SMT - Modeling

• **Modeling Pr(elf):** describing the relevant dependencies between e and f

• Here: log-linear model that combines feature functions depending on Chinese string (f), English string (e) (+ word alignment)

\[ \Pr(e|f) = p_{\lambda_1}^M(e|f) \propto \exp\left[ \sum_{m=1}^{M} \lambda_m h_m(e, f) \right] \]
SMT - Training

- **Training** of model parameters
- Here: maximum BLEU training
  - Discriminative training with BLEU as objective function

\[ \hat{\lambda}_1^M = \arg\max \text{BLEU}(r_1^S, \hat{e}(f_1^S; \lambda_1^M)) \]

- Algorithm: greedy descent with optimal line search
- Advantage: directly optimizes evaluation criterion
- Problem: danger of overfitting to training data
SMT - Search

- Search: minimize expected loss
- Standard approach: 0-1 loss function
  - Later: Minimum Bayes risk with different loss functions (e.g. BLEU loss(!))
- Log-linear model: simple decision rule

\[ \hat{e}(f) = \arg\max_e \left\{ \sum_{m=1}^{M} \lambda_m h_m(f, e) \right\} \]
Baseline: Alignment Templates

Basic idea: learn all aligned phrase pairs (= Alignment Templates) seen in training data
Baseline: Features

• Product of alignment template probabilities
• Product of word translation probabilities
• (4) Trigram language models
• Number of produced words
• Number of produced alignment templates
• … Some other feature functions
The Plan

- Starting Point: state-of-the-art phrase-based statistical MT system
  - Here: Alignment Template system from ISI
  - Purely data-driven
- Error Analysis: What goes wrong?
- Develop syntactically motivated feature functions for specific problems
The Plan - Feature Functions

- Refined feature functions depend on:
  - Standard: English string, Chinese string, word alignment, phrase alignment
  - POS tag sequence
  - Chunk segmentation for Chinese/English
  - Parse trees for Chinese/English
  - Dependency parses for Chinese/English
  - ...
The Plan - Major Tasks

1. Error Analysis/Feature Hunting
   • Contrastive error analysis

2. Development of Feature Functions

3. Discriminative Training Techniques
   • Maximum-BLEU, Perceptron

4. Search Approaches
   • Minimum Bayes risk with syntactic loss functions
The Plan - Framework (1)

- **Chinese-English large data track**
  - Training data for Training of Baseline FF
    - 150M words per language
  - Chinese treebank available
  - **Dev** data for Maximum BLEU training
    - NIST eval-01: 993 sentences
  - **Test** data
    - NIST eval-02: 878 sentences
  - Blind test data: (for after-workshop evaluation)
    - NIST eval-03: 929 sentences
  - Prepared also larger sets of development/test data
    - Dev: 4830 sentences, Test: 1813 sentences
During workshop: Rescoring of n-best lists

- Advantages:
  - No need to integrate FF in dynamic programming search
  - FF can depend arbitrarily on full Chinese/English sentence/parsing tree/…
  - Simple software architecture
  - Using fixed set of n-best alternatives: FF = vector of numbers

- (Disadvantage: limits improvements to n-best lists)

- N=16384 alternative translations for Dev/Test
  - Precomputed before workshop
Evaluation metric: BLEU

- Automatic evaluation via BLEU
  - Compute precision of n-grams (here: 1,2,3,4-grams)
  - Geometric mean of n-gram precision
  - Multiply a penalty factor for too short translations
- Shown in recent evaluations to correlate with humans
  - But: MT performance looks too good compared to human performance measured by BLEU
- Important question for workshop: Is BLEU sensitive to (subtle) syntactic changes?
- After-workshop: subjective evaluation of final system
Oracle BLEU

• How large are potential improvements?
  – Oracle(-Best) translations: set of translations from n-best list that give best score
  – Oracle-Worst translations: set of worst scoring translations

• Computation for BLEU non-trivial
  – Greedy search algorithm
Oracle vs. anti-Oracle

The graph shows the comparison between Oracle BLEU and anti-Oracle BLEU across different data sizes. The x-axis represents data sizes in powers of 2 (1, 4, 16, 64, 256, 1024, 4096, 16384), and the y-axis represents BLEU scores in percentages. There are two lines: one for Oracle BLEU (red) and one for anti-Oracle BLEU (yellow). As the data size increases, the Oracle BLEU score increases, while the anti-Oracle BLEU score decreases.
How good is the oracle?

- Average human score
  - BLEU: 37.9% (3 refs)
- Average first-best score
  - BLEU: 27.6% (3 refs)
  - Relative to human BLEU score: 72.9%
- Average oracle of 1000 best BLEU score
  - BLEU: 39.8% (3 refs)
  - Relative to human BLEU score: 105.0%
- But: quality of oracle is still bad…
  - Hence: reaching oracle is (very) unrealistic
  - Important to note: references used were taken into account during selection process of oracle
  - This experiment does not show that BLEU is not a good measure to assess MT quality
Processing Noisy Data

Tagger tries to “fix up” ungrammatical sentences:

China NNP 14 CD open JJ border NN cities NNS
achievements VBZ remarkable JJ

MT data include headlines with no verb.

Tagger trained on full sentences with normalized punctuation:

China NNP ’s POS economic JJ development NN and CC
opening VBG up RP 14 CD border NN cities NNS
remarkable JJ achievements .
Processing Noisy Data

Parser can create verb phrases where none exist:

```
S
├── NP
│   ├── NNP: China
│   ├── CD: 14
│   ├── JJ: open
│   ├── NN: border
│   ├── NNS: cities
│   └── VP
│       ├── NN: construction
│       └── NP
│           ├── JJ: remarkable
│           └── NNS: achievements
```

Syntax for Statistical MT JHU 2003 WS
Processing Noisy Data

Implications

- Features such “is there a verb phrase” may not do what you expect
- Possible solution: features involving probabilities of parse/tag sequence - “how good a verb phrase?”
Dependency Trees: Feature Extraction

- **Idea**
  - Use dependency trees to eliminate non-terminal super-structure above lexical items
  - Represent more directly the relationships between words
Aligned Dependency Tree Features

- Modify regular tree-to-tree alignment to align dependency trees
  - Consider translation occurring at every node
  - Introduce lexical reordering probability

- Train aligner on about 40,000 pairs of FBIS Chinese/English sentences converted to dependency trees

- Extract features from alignments
  - Alignment score
  - Number of similar nodes (e.g. verbs) that align
Sample Alignment

\[ h_{\text{ALIGNSCORE}}(e, f) = \text{alignment probability (3.940690245314e-26)} \]

\[ h_{\text{VERBALIGN}}(e, f) = \text{Number of times a verb aligns with another verb (0)} \]
Translation Results for Dependency Derived Features

Investigated many variants on dependency-derived features. Combined them one-at-a-time with features from the baseline system (baseline: 31.6%).

\[ P(e | f) \propto \exp \left\{ \sum_{m=1}^{M} \lambda_m h_m(e, f) + \lambda_{M+1} h_{M+1}(e, f) \right\} \]

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both parses have verbs</td>
<td>31.6%</td>
</tr>
<tr>
<td>Verb Alignment</td>
<td>31.6%</td>
</tr>
<tr>
<td>Difference in # Nouns</td>
<td>31.4%</td>
</tr>
<tr>
<td>Dependency Projection</td>
<td>31.6%</td>
</tr>
<tr>
<td>Doubly Transitive Dependencies</td>
<td>31.4%</td>
</tr>
<tr>
<td>Maximum LM Ratio</td>
<td>31.6%</td>
</tr>
<tr>
<td>Average LM Ratio</td>
<td>31.3%</td>
</tr>
<tr>
<td>Minimum LM Ratio</td>
<td>31.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of args of main verb</td>
<td>31.6%</td>
</tr>
<tr>
<td>Noun Alignment</td>
<td>31.4%</td>
</tr>
<tr>
<td>Difference in # Verbs</td>
<td>31.4%</td>
</tr>
<tr>
<td>Transitive Eng Dependencies</td>
<td>31.7%</td>
</tr>
<tr>
<td>Labeled-Dep Model-1 Prob</td>
<td>31.5%</td>
</tr>
<tr>
<td>Alignment Score</td>
<td>31.6%</td>
</tr>
<tr>
<td>Dependency LM Score</td>
<td>31.7%</td>
</tr>
<tr>
<td>Dependency LM Ratio</td>
<td>31.5%</td>
</tr>
</tbody>
</table>