An Empirical Comparison of Features and Tuning for Phrase-based Machine Translation

Spence Green

with Daniel Cer and Chris Manning

Stanford University

WMT // 27 June 2014
Recap: ACL'13 Results

SGD-based, $n$-best learning \hspace{2cm} L_1 \text{ feature selection}
Recap: ACL13 Results

SGD-based, $n$-best learning

$L_1$ feature selection

**BOLT-scale** Zh–En on NIST data:

<table>
<thead>
<tr>
<th>BLEU</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERT</td>
<td>48.4</td>
</tr>
</tbody>
</table>
Recap: ACL13 Results

SGD-based, \( n \)-best learning \quad L_1 \text{ feature selection}

**BOLT-scale** Zh–En on NIST data:

<table>
<thead>
<tr>
<th>BLEU</th>
<th>( \Delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERT</td>
<td>48.4</td>
</tr>
<tr>
<td>SGD</td>
<td>48.1</td>
</tr>
</tbody>
</table>
Recap: ACL13 Results

SGD-based, \( n \)-best learning \quad L_1 \) feature selection

BOLT-scale Zh–En on NIST data:

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>( \Delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MERT</td>
<td>48.4</td>
<td></td>
</tr>
<tr>
<td>SGD</td>
<td>48.1</td>
<td></td>
</tr>
<tr>
<td>SGD+Features</td>
<td>49.9</td>
<td>+1.5 : - )</td>
</tr>
</tbody>
</table>
Motivation #1: WMT13 Shared Task :-(

![Graph showing BLEU scores for dense and feature-rich models over epochs.](image-url)
Motivation #1: WMT13 Shared Task

En–Fr news2012 (dev)

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>31.1</td>
</tr>
<tr>
<td>SGD+Features</td>
<td>31.5  +0.4</td>
</tr>
</tbody>
</table>
Motivation #2: Practical Issues

Q1: Which **phrase-based features** should I use?
Motivation #2: Practical Issues

Q1: Which phrase-based features should I use?

Q2: Why don’t my features help?
My Frustrating Summer...

What’s wrong with feature-rich MT?

1. Loss Function
My Frustrating Summer...

What’s wrong with feature-rich MT?

1. Loss Function
2. References and scoring functions
My Frustrating Summer...

What’s wrong with feature-rich MT?

1. Loss Function
2. References and scoring functions
3. Representation: Features
What’s wrong with feature-rich MT?

1. Loss Function
2. References and scoring functions
3. Representation: **Features**

This paper as a pain reliever...
Loss Function
ACL13: Online PRO

Sensitive to length

Doesn’t optimize top-$k$

Slow to compute (sampling)
This work: Online Expected Error

Expected BLEU

\[ \ell_t(w_{t-1}) = E_{p_{w_{t-1}}} [-\text{BLEU}(d)] \]

\[ = - \sum_{d \in H} p_{w_{t-1}}(d) \cdot \text{BLEU}(d) \]
This work: Online Expected Error

Expected BLEU

\[ l_t(w_{t-1}) = E_{p_{w_{t-1}}} [-BLEU(d)] \]

\[ = - \sum_{d \in H} p_{w_{t-1}}(d) \cdot BLEU(d) \]

Smooth, non-convex

Fast, less sensitive to length

...but still doesn’t prefer top-$k$
References and Scoring
Single vs. Multiple References

**Experiment**: Compute BLEU+1 for each reference
Single vs. Multiple References

Experiment: Compute BLEU+1 for each reference

Baseline MT system
Single vs. Multiple References

**Experiment:** Compute BLEU+1 for each reference

Baseline MT system

Ar–En NIST MT05 has five (5) references
MT05: Max. vs. Min. BLEU+1
MT05: Max. vs. All References BLEU+1
Refs and Scoring Functions

Single-ref Lesson: Don’t try too hard
Refs and Scoring Functions

**Single-ref Lesson**: Don’t try too hard

Blame the *scoring function*?

- BLEU+1
- BLEU-Nakov [Nakov et al. 2012]
Refs and Scoring Functions

**Single-ref Lesson:** Don’t try too hard

Blame the **scoring function?**

- **BLEU+1**
- **BLEU-Nakov**  [Nakov et al. 2012]
- **BLEU+Noise**  add Gaussian noise to $n$-gram precisions
Refs and Scoring Functions

**Single-ref Lesson**: Don’t try too hard

Blame the **scoring function**?

- BLEU+1
- BLEU-Nakov $^*$ [Nakov et al. 2012]
- BLEU+Noise: add Gaussian noise to $n$-gram precisions
- TER (short translations)
- Linear combinations
Representation: Features
Representation: Dense + Extended

Dense features

Moses baseline templates [Koehn et al. 2007]
Hierarchical lex. reordering [Galley and Manning 2008]
Rule count and uniqueness indicator
Representation: Dense + Extended

Dense features

Moses baseline templates [Koehn et al. 2007]
Hierarchical lex. reordering [Galley and Manning 2008]
Rule count and uniqueness indicator

Extended features
Representation: Dense + Extended

Dense features

Moses baseline templates [Koehn et al. 2007]
Hierarchical lex. reordering [Galley and Manning 2008]
Rule count and uniqueness indicator

Extended features

Fire less than dense but more than sparse
**Goal**: a general, robust feature-rich model
**Goal**: a general, robust feature-rich model

No more ad-hoc features
**Goal:** a general, robust feature-rich model

No more ad-hoc features

Starting point for more specific features
Five Feature Categories

Common MT error types

1. Lexical Choice
2. Word Alignments
3. Phrase Boundaries
4. Derivation Quality
5. Reordering
Five Feature Categories

Common MT error types

1. Lexical Choice
2. Word Alignments
3. Phrase Boundaries
4. Derivation Quality
5. Reordering

Sources: Novel, literature, word-of-mouth, etc.
Features: Lexical Choice

**Filtered** Rule Indicator

\[
\begin{array}{c|c}
\text{maison} & \text{maison->the\_house} \\
\text{the house} &
\end{array}
\]
Features: Lexical Choice

**Filtered** Rule Indicator

- *maison* → *the_house*
- the house

**Class-based variant**

- *maison* → 64->22_14
- the house
Features: Lexical Choice

Target unigram class

e: utility stocks lead shares higher
Features: Lexical Choice

Target unigram class

e: utility stocks lead shares higher

77 82 3 82 267
Features: Lexical Choice

Target unigram class

e: utility stocks lead shares higher

77 82 3 82 267

Feature strings:
CLASS: 77
CLASS: 82
CLASS: 3
CLASS: 82
CLASS: 82
CLASS: 267
Features: Word Alignments

- **tarceva**
- **parvient**
- **ainsi**
- **à**
- **stopper**
- **la**
- **croissance**

Feature strings:
- **ALGN:** parvient -> able
- **ALGN:** stopper -> to_halt

Example:
- tarceva was thus able to halt the growth
Features: Word Alignments

Feature strings:

ALGN:parvient->able
ALGN:stopper->to_halt
etc.
Features: Phrase Boundaries

Target bigram phrase boundary

e: utility | stocks lead shares | higher
Features: Phrase Boundaries

Target bigram phrase boundary

\[ e: \text{utility} \mid \text{stocks} \mid \text{lead} \mid \text{shares} \mid \text{higher} \]

\[ 77 \mid 82 \mid 3 \mid 82 \mid 267 \]
Features: Phrase Boundaries

Target bigram phrase boundary

<table>
<thead>
<tr>
<th>e: utility</th>
<th>stocks</th>
<th>lead</th>
<th>shares</th>
<th>higher</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>82</td>
<td>3</td>
<td>82</td>
<td>267</td>
</tr>
</tbody>
</table>

Feature strings:

BOUNDARY: 77_82
BOUNDARY: 82_267
Features: Derivation Quality

Rule dimension features

\[ \textit{maison} \Rightarrow \text{the house} \]
Features: Derivation Quality

Rule dimension features

\[maison \Rightarrow \text{the house}\]

Feature strings:

\[
\begin{align*}
\text{SOURCE\_DIM} & : 1 \\
\text{TARGET\_DIM} & : 2 \\
\text{DIM} & : 1-2
\end{align*}
\]
Features: Reordering

**Filtered** Rule Orientation

```
maison           SWAP:maison->the_house
the house
```
Features: Reordering

Filtered Rule Orientation

\[ \text{maison} \quad | \quad \text{SWAP: maison} \rightarrow \text{the house} \]
\[ \text{the house} \]

Class-based variant

\[ \text{maison} \quad | \quad \text{SWAP: 64} \rightarrow 22_{-14} \]
\[ \text{the house} \]
Aside: Learning Word Classes

**Experiment**: 3.7M English tokens, 512 classes

[Whittaker and Woodland 2001]
[Uszkoreit and Brants 2008]
Aside: Learning Word Classes

**Experiment:** 3.7M English tokens, 512 classes

<table>
<thead>
<tr>
<th></th>
<th>#threads</th>
<th>min:sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown (wcluster)</td>
<td>1</td>
<td>1023:39</td>
</tr>
</tbody>
</table>

[Whittaker and Woodland 2001](#)

[Uszkoreit and Brants 2008](#)
Aside: Learning Word Classes

**Experiment:** 3.7M English tokens, 512 classes

<table>
<thead>
<tr>
<th></th>
<th>#threads</th>
<th>min:sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown (wcluster)</td>
<td>1</td>
<td>1023:39</td>
</tr>
<tr>
<td>Clark (cluster_neyessen)</td>
<td>1</td>
<td>890:11</td>
</tr>
</tbody>
</table>
Aside: Learning Word Classes

**Experiment:** 3.7M English tokens, 512 classes

<table>
<thead>
<tr>
<th>Method</th>
<th>#threads</th>
<th>min:sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown (wcluster)</td>
<td>1</td>
<td>1023:39</td>
</tr>
<tr>
<td>Clark (cluster_neyessen)</td>
<td>1</td>
<td>890:11</td>
</tr>
<tr>
<td>Och (mkcls)</td>
<td>1</td>
<td>199:04</td>
</tr>
</tbody>
</table>

[Whittaker and Woodland 2001](#)

[Uszkoreit and Brants 2008](#)
Aside: Learning Word Classes

**Experiment:** 3.7M English tokens, 512 classes

<table>
<thead>
<tr>
<th></th>
<th>#threads</th>
<th>min:sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown (wcluster)</td>
<td>1</td>
<td>1023:39</td>
</tr>
<tr>
<td>Clark (cluster_neyessen)</td>
<td>1</td>
<td>890:11</td>
</tr>
<tr>
<td>Och (mkcls)</td>
<td>1</td>
<td>199:04</td>
</tr>
<tr>
<td><strong>This paper</strong></td>
<td>8</td>
<td><strong>2:42</strong></td>
</tr>
</tbody>
</table>

[Whittaker and Woodland 2001][Uszkoreit and Brants 2008]
Experiments
NIST Experiments

**Stanford Phrasal**

**BOLT-scale systems**: Ar–En, Zh–En

Four references, uncased BLEU-4

[Green et al. 2014]
### NIST Results: Ar–En

<table>
<thead>
<tr>
<th>Dense</th>
<th>BLEU</th>
<th>(\Delta)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>42.2</td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>BLEU</td>
<td>Δ</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Dense</td>
<td>42.2</td>
<td></td>
</tr>
<tr>
<td>Dense + ACL13</td>
<td>44.0</td>
<td>+1.8</td>
</tr>
</tbody>
</table>

NIST Results: Ar–En

Zh–En: +2.0 BLEU
<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>42.2</td>
<td></td>
</tr>
<tr>
<td>Dense + ACL13</td>
<td>44.0</td>
<td>+1.8</td>
</tr>
<tr>
<td>Dense + Ext</td>
<td>44.6</td>
<td>+2.4</td>
</tr>
</tbody>
</table>
## NIST Results: Ar–En

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>42.2</td>
<td></td>
</tr>
<tr>
<td>Dense + ACL13</td>
<td>44.0</td>
<td>+1.8</td>
</tr>
<tr>
<td>Dense + Ext</td>
<td>44.6</td>
<td>+2.4</td>
</tr>
<tr>
<td>Dense + Ext + <strong>Domain</strong></td>
<td>45.0</td>
<td>+2.8</td>
</tr>
</tbody>
</table>

**Domain**: feature space augmentation

[Daumé III 2007]
NIST Results: Ar–En

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>42.2</td>
<td></td>
</tr>
<tr>
<td>Dense+ACL13</td>
<td>44.0</td>
<td>+1.8</td>
</tr>
<tr>
<td>Dense+Ext</td>
<td>44.6</td>
<td>+2.4</td>
</tr>
<tr>
<td>Dense+Ext+<strong>Domain</strong></td>
<td>45.0</td>
<td>+2.8</td>
</tr>
</tbody>
</table>

**Domain**: feature space augmentation [Daumé III 2007]

Zh–En: +2.0 BLEU
WMT–14 Shared Task

**Single reference**, uncased BLEU-4
WMT–14 Shared Task

**Single reference**, uncased BLEU-4

**All Fr–En constrained data**

<table>
<thead>
<tr>
<th>Bilingual</th>
<th>Monolingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Segments</td>
<td>#Tokens</td>
</tr>
<tr>
<td>36.3M</td>
<td>2.1M</td>
</tr>
<tr>
<td></td>
<td>7.2B</td>
</tr>
</tbody>
</table>
## WMT–14 Results: Fr–En

<table>
<thead>
<tr>
<th>Model</th>
<th>2014 BLEU</th>
<th>(\Delta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>35.6</td>
<td></td>
</tr>
<tr>
<td>Dense + Ext</td>
<td>36.0</td>
<td>+0.4</td>
</tr>
</tbody>
</table>

Uncased BLEU: 1st place
Manual eval: 2–4 cluster
### WMT–14 Results: Fr–En

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU Score</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>35.6</td>
<td></td>
</tr>
<tr>
<td>Dense+Ext</td>
<td>36.0</td>
<td>+0.4</td>
</tr>
</tbody>
</table>

Uncased BLEU: 1st place

Manual eval: 2–4 cluster
## Analysis: Single vs. Multiple References

### Ar–En MT09 results

<table>
<thead>
<tr>
<th></th>
<th>4-ref</th>
<th>Δ</th>
<th>1-ref</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>48.0</td>
<td></td>
<td>47.8</td>
<td></td>
</tr>
<tr>
<td>Dense + Ext</td>
<td>50.0</td>
<td>+2.0</td>
<td>48.9</td>
<td>+1.1</td>
</tr>
</tbody>
</table>
General Observations

More expressive models match refs better (duh)

Single-ref condition == overfitting
General Observations

More expressive models match refs better (duh)

Single-ref condition == overfitting

Sensitivity to tuning set size/content

Bitext tuning
General Observations

More expressive models match refs better (duh)

Single-ref condition == overfitting

Sensitivity to tuning set size/content

Bitext tuning

Ablation isn’t very helpful

Approximate search, non-convex
Conclusion and Impact

Baseline feature-rich representation

Domain adaptation
Conclusion and Impact

Baseline feature-rich representation

Domain adaptation

Faster, better online tuning
Conclusion and Impact

**Baseline feature-rich representation**

Domain adaptation

Faster, better online tuning

Scalable software to implement the features

**See new Phrasal release**
An Empirical Comparison of Features and Tuning for Phrase-based Machine Translation

Spence Green

with Daniel Cer and Chris Manning

Stanford University

WMT // 27 June 2014