Adaptive MT Systems for Post-Editing Tasks
Machine Translation for Human Translators

Michael Denkowski, Alon Lavie, Chris Dyer, Jaime Carbonell, Gregory Shreve*, Isabel Lacruz*

Carnegie Mellon University, *Kent State University

May 13, 2015
When is a translation “good enough”? 
Total solar eclipse spectacle with fire Arctic tour

2015-03-23 07:31 Source: People’s Daily

UK GMT at 10:11 on March 20, a rare solar eclipse spectacle will come to Europe. According to local media reports, Norway, under the 1300 km distance Arctic Norway Svalbard and membership Danish Faroe Islands are the two can watch the full eclipse of the best observation points. In most European countries, the observer can only see part of the sun is blocked by the shadow of the moon, that is, the partial eclipse of the landscape.

This is the 1954 total solar eclipse once again usher in mainland Norway. According to the Norwegian local astronomers predicted from the next total solar eclipse occurred on Norwegian time to wait for 46 years, that is 2061. The next solar eclipse occurs recent times and the country was March 9, 2016 Sumatra; August 21, 2017 the United States and Chile, July 2, 2019 in.
4 **daily operations**

**Practice**

1. to maintain proper startup speed, not too much.

2. to observe the *voltmeter* voltage, if the voltage is below the limit, the truck should be stopped immediately.

3. truck during walking, but not allowed to flip the direction switch to change the direction of travel, to prevent burn damage electrical components and gear.

4. travel and promotion should not be carried out simultaneously.

5. Note the drive system, steering system is sound normal, abnormal sound should be immediately removed *fault*, prohibited ill job.

6. when the transition to slow down in advance.

7. when the poor road conditions in the job, it is important to reduce the appropriate and should reduce speed.
Human Translation

International Organizations

Global Businesses

Community Projects

Require human-quality translation of complex content

Machine translation currently unable to deliver quality and consistency
Human Translation

International Organizations

Global Businesses

Community Projects

$37 billion in 2014

Require human-quality translation of complex content

Machine translation currently unable to deliver quality and consistency
Use **machine translation** to improve speed of **human translation**
MT with Human Post-Editing

Source Document → Machine Translation → Human Editing → Translated Document

Use **machine translation** to improve speed of **human translation**

Increasing adoption by **government organizations and businesses**
Son comportement ne peut être qualifié que d’irréprochable.
Son comportement ne peut être qualifié que d’irréprochable.

His behavior cannot be described as d’irréprochable.
Son comportement ne peut être qualifié que d’irréprochable.

His behavior cannot be described as d’irréprochable.

Its behavior can only be described as flawless.
Son comportement ne peut être qualifié que d’irréprochable.

His behavior cannot be described as d’irréprochable.

Its behavior can only be described as flawless.

MT task: minimize work for human translators
Post-editing faster and more accurate than unaided translation (Guerberof, 2009; Carl et al., 2011; Koehn, 2012; Zhechev, 2012; inter alia)

Productivity gains but MT systems not engineered for human post-editing

How can we extend MT systems to target post-editing?
Overview

Online learning for statistical MT
- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments
- Kent State live post-editing

Automatic metrics for post-editing
- Meteor automatic metric
- Evaluation and optimization for post-editing

Conclusion
Overview

Online learning for statistical MT
- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments
- Kent State live post-editing

Automatic metrics for post-editing
- Meteor automatic metric
- Evaluation and optimization for post-editing

Conclusion
Online Learning for MT

Statistical translation models built from bilingual data
Online Learning for MT

Statistical translation models built from bilingual data

Post-editing generates new bilingual data
Online Learning for MT

Statistical translation models built from bilingual data

Post-editing generates new bilingual data

Goal: incorporate post-editing data back into model in real time

Learn from feedback: avoid repeating the same translation errors
Online Learning for MT

Batch learning (standard MT):

Estimation → Prediction

Requirement: all system components operate at the sentence level
Online Learning for MT

Batch learning (standard MT):

Estimation $\rightarrow$ Prediction

Online learning (this work):

Prediction $\rightarrow$ Truth $\rightarrow$ Update

Requirement: all system components operate at the sentence level
Online Learning for MT

Batch learning (standard MT):

Estimation → Prediction

Online learning (this work):

Prediction → Truth → Update

Requirement: all system components operate at the sentence level
Overview

Online learning for statistical MT

- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments

- Kent State live post-editing

Automatic metrics for post-editing

- Meteor automatic metric
- Evaluation and optimization for post-editing

Conclusion
Post-Editing with Standard MT

Static

Large LM

Grammar

Weights

Input Sentence

Decoder

Post-Editing

$X \rightarrow f/e$

$w_i \ldots w_n$
Phrase-based machine translation (Koehn et al., 2003):

\[ \text{la vérité} \rightarrow \text{the truth} \]

- Match spans of input text against phrases we know how to translate
Phrase-based machine translation (Koehn et al., 2003):

\[ \text{la vérité} \rightarrow \text{the truth} \]

- Match spans of input text against phrases we know how to translate

Hierarchical phrase-based MT (Chiang, 2007):

\[ X \rightarrow \text{la } X_{1} \text{ / the } X_{1} \]
\[ X \rightarrow \text{vérité / truth} \]

- Generalization where phrases can contain other phrases
- Phrases become rules in a synchronous context-free grammar
Hierarchical Phrase-Based Translation Example

Input sentence: Pourtant, la vérité est ailleurs selon moi.

Translation Grammar:

\[ X \rightarrow X_1 \text{ est ailleurs } X_2 \ . \ / \ X_2 \ , \ X_1 \text{ lies elsewhere} \ . \]
\[ X \rightarrow \text{Pourant} \ , \ / \text{Yet} \]
\[ X \rightarrow \text{la vérité} \ / \text{the truth} \]
\[ X \rightarrow \text{selon moi} \ / \text{in my view} \]

Glue Grammar:

\[ S \rightarrow S_1 \ X_2 \ / \ S_1 \ X_2 \]
\[ S \rightarrow X_1 \ / \ X_2 \]
Pourtant, la vérité est ailleurs selon moi.
Pourtant, la vérité est ailleurs selon moi.

the truth
Hierarchical Phrase-Based Translation Example

Pourant, la vérité est ailleurs selon moi.

Yet in my view the truth
Hierarchical Phrase-Based Translation Example

Pourant, la vérité est ailleurs selon moi.

Yet, the truth lies elsewhere.

Pourant, X estailleurs X.

Yet, X, X lies elsewhere.

la vérité
selon moi

in my view
the truth

F

E
Ambiguity: many ways to translate the same source phrase

Add **feature scores** that encode properties of translation:

\[
\begin{align*}
X & \rightarrow \text{devis / quote} & 0.5 & 10 & -137 & \ldots \\
X & \rightarrow \text{devis / estimate} & 0.4 & 13 & -261 & \ldots \\
X & \rightarrow \text{devis / specifications} & 0.2 & 5 & -407 & \ldots
\end{align*}
\]

Decoder uses **feature scores** and **weights** to select the most likely translation derivation.
Linear Translation Models

Single feature score for a translation derivation with rule-local features $h_i \in H_i$:

$$H_i(D) = \sum_{X \rightarrow \tilde{f}/\tilde{e} \in D} h_i (X \rightarrow \tilde{f}/\tilde{e})$$

Score for a derivation using several features $H_i \in H$ with weight vector $w_i \in W$:

$$S(D) = \sum_{i=1}^{\lvert H \rvert} w_i H_i(D)$$

Decoder selects translation with largest product $W \cdot H$
Linear Translation Models

Single feature score for a translation derivation with rule-local features $h_i \in H_i$:

$$H_i(D) = \sum_{X \rightarrow \overline{f}/\overline{e} \in D} h_i(X \rightarrow \overline{f}/\overline{e})$$

Score for a derivation using several features $H_i \in H$ with weight vector $w_i \in W$:

$$S(D) = \sum_{i=1}^{\left|H\right|} w_i H_i(D)$$

Decoder selects translation with largest product $W \cdot H$

✓ sentence-level prediction step
Learning translations
Devis de garage en quatre étapes. Avec l’outil Auda-Taller, l’entreprise Audatex garantit que l’usager obtient un devis en seulement quatre étapes : identifier le véhicule, chercher la pièce de rechange, créer un devis et le générer. La facilité d’utilisation est un élément essentiel de ces systèmes, surtout pour convaincre les professionnels les plus âgés qui, dans une plus ou moins grande mesure, sont rétifs à l’utilisation de nouvelles techniques de gestion.

A shop’s estimate in four steps. With the AudaTaller tool, Audatex guarantees that the user gets an estimate in only 4 steps: identify the vehicle, look for the spare part, create an estimate and generate an estimate. User friendliness is an essential condition for these systems, especially to convincing older technicians, who, to varying degrees, are usually more reluctant to use new management techniques.
Devis de garage en quatre étapes. Avec l’outil Auda-Taller, l’entreprise Audatex garantit que l’usager obtient un devis en seulement quatre étapes : identifier le véhicule, chercher la pièce de rechange, créer un devis et le générer. La facilité d’utilisation est un élément essentiel de ces systèmes, surtout pour convaincre les professionnels les plus âgés qui, dans une plus ou moins grande mesure, sont rétifs à l’utilisation de nouvelles techniques de gestion.

A shop’s estimate in four steps. With the AudaTaller tool, Audatex guarantees that the user gets an estimate in only 4 steps: identify the vehicle, look for the spare part, create an estimate and generate an estimate. User friendliness is an essential condition for these systems, especially to convincing older technicians, who, to varying degrees, are usually more reluctant to use new management techniques.
Model Estimation: Word Alignment
Brown et al. (1993), Dyer et al. (2013)

F

Devis de garage en quatre étapes

A shop’s estimate in four steps

E
Model Estimation: Word Alignment
Brown et al. (1993), Dyer et al. (2013)

Devis de garage en quatre étapes
A shop's estimate in four steps
## Model Estimation: Phrase Extraction

Koehn et al. (2003), Och and Ney (2004), Och et al. (1999)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>shop</th>
<th>'s</th>
<th>estimate</th>
<th>in</th>
<th>four</th>
<th>steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>garage</td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>quatre</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>étapes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
</tbody>
</table>
### Model Estimation: Phrase Extraction

Koehn et al. (2003), Och and Ney (2004), Och et al. (1999)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>shop</th>
<th>'s</th>
<th>estimate</th>
<th>in</th>
<th>four</th>
<th>steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>●</td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>garage</td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>quatre</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>étapes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
</tbody>
</table>

- de garage \(\downarrow\) a shop’s
- en quatre étapes \(\downarrow\) in four steps
Yet, in my view, the truth lies elsewhere.

Pourtant, selon moi, la vérité est ailleurs.

✓ sentence-level rule learning
Chiang (2007)

<table>
<thead>
<tr>
<th></th>
<th>Yet</th>
<th>in</th>
<th>my</th>
<th>view</th>
<th>the</th>
<th>truth</th>
<th>lies</th>
<th>elsewhere</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pourtant</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>,</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>la</td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vérité</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>est</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ailleurs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>selon</td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>moi</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
<td></td>
</tr>
</tbody>
</table>

la vérité est ailleurs selon moi. ➞ in my view, the truth lies elsewhere.
la vérité est ailleurs selon moi . $\rightarrow$ in my view , the truth lies elsewhere .

$X_1$ est ailleurs $X_2$ . $\rightarrow$ $X_2$ , $X_1$ lies elsewhere .
✓ sentence-level rule learning

\[
\begin{array}{cccccccc}
 & \text{Yet} & \text{in} & \text{my} & \text{view} & \text{the} & \text{truth} & \text{lies} & \text{elsewhere} \\
\hline
\text{Pourtant} & \bullet & & & & & & \\
, & & & & & & & \\
\text{la} & & & & & \bullet & & \\
\text{vérité} & & & & \bullet & & & \\
\text{est} & & & & & & \bullet & \\
\text{ailleurs} & & & & & \bullet & & \\
\text{selon} & \bullet & \bullet & & & & & \\
\text{moi} & & & & \bullet & & & \\
. & & & & & & & \\
\end{array}
\]

\[X^1_1\] est ailleurs selon moi . \[\rightarrow\] in my view , the truth lies elsewhere .

\[X^1_1\] est ailleurs \[X^2_2\] . \[\rightarrow\] \[X^2_2\] , \[X^1_1\] lies elsewhere .
Parameterization: Feature Scoring

Add *feature functions* to rules $X \rightarrow \bar{f}/\bar{e}$:

- Training Data
- Corpus Stats
- Scored Grammar (Global)

Translate Sentence

Input Sentence

Static
Parameterization: Feature Scoring

Add feature functions to rules $X \rightarrow \bar{f}/\bar{e}$:

- Training Data
- Corpus Stats
- Scored Grammar (Global)
- Translate Sentence

$\sum_{i=1}^{N}$

Static

Input Sentence

$X \rightarrow \bar{f}/\bar{e}$

$\times$ corpus-level rule scoring
Suffix Array Grammar Extraction

Static

Training Data → Suffix Array

SA Sample → Sample Stats → Grammar (Sentence) → Translate Sentence

Input Sentence

\[ \sum_{i=1}^{N} \]

\[ X \rightarrow f/e \]
Scoring via Sampling

Suffix array statistics available in sample $S$ for each source $\bar{t}$:

- $c_S(\bar{f}, \bar{e})$: count of instances where $\bar{f}$ is aligned to $\bar{e}$ (co-occurrence count)
- $c_S(\bar{f})$: count of instances where $\bar{f}$ is aligned to any target
- $|S|$: total number of instances (equal to occurrences of $\bar{f}$ in training data, up to the sample size)

Used to calculate feature scores for each rule at the time of extraction
Scoring via Sampling

Suffix array statistics available in sample \( S \) for each source \( \bar{f} \):

- \( c_S(\bar{f}, \bar{e}) \): count of instances where \( \bar{f} \) is aligned to \( \bar{e} \)
  (co-occurrence count)

- \( c_S(\bar{f}) \): count of instances where \( \bar{f} \) is aligned to any target

- \( |S| \): total number of instances
  (equal to occurrences of \( \bar{f} \) in training data, up to the sample size)

Used to calculate feature scores for each rule at the time of extraction

× sentence-level grammar extraction, but static training data
Overview

Online learning for statistical MT

- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments

- Kent State live post-editing

Automatic metrics for post-editing

- Meteor automatic metric
- Evaluation and optimization for post-editing

Conclusion
Online Grammar Extraction
Denkowski et al. (EACL 2014)

Static

Training Data → Suffix Array

Sample → \( \sum_{i=1}^{N} \) Sample Stats → Grammar (Sentence) → Translate Sentence

Input Sentence

\( X \rightarrow \bar{f}/\bar{e} \)
Online Grammar Extraction
Denkowski et al. (EACL 2014)

Static

Training Data → Suffix Array

Sample Stats

Dynamic

Lookup Table

Post-Edit Sentence

Input Sentence

Translate Sentence

Grammar (Sentence)

\[ \sum_{i=1}^{N} \]

\[ X \rightarrow \bar{f}/\bar{e} \]
Maintain dynamic **lookup table** for post-edit data

Pair each sample $S$ from suffix array with **exhaustive** lookup $L$ from lookup table

Parallel statistics available at grammar scoring time:

- $c_L(\bar{f}, \bar{e})$: count of instances where $\bar{f}$ is aligned to $\bar{e}$ (co-occurrence count)

- $c_L(\bar{f})$: count of instances where $\bar{f}$ is aligned to any target

- $|L|$: total number of instances (equal to occurrences of $\bar{f}$ in post-edit data, no limit)
Suffix array feature set (Lopez 2008)

Phrase features encode likelihood of translation rule given training data

Features scored with $S$:

\[
\text{CoherentP}(e|f) = \frac{c_S(f, e)}{|S|}
\]

\[
\text{Count}(f, e) = c_S(f, e)
\]

\[
\text{SampleCount}(f) = |S|
\]
Rule Scoring
Denkowski et al. (EACL 2014)

Suffix array feature set (Lopez 2008)

Phrase features encode likelihood of translation rule given training data

Features scored with $S$ and $L$:

$$\text{CoherentP}(e|f) = \frac{c_S(f, e) + c_L(f, e)}{|S| + |L|}$$

$$\text{Count}(f, e) = c_S(f, e) + c_L(f, e)$$

$$\text{SampleCount}(f) = |S| + |L|$$
Indicator features identify certain classes of rules

Features scored with S:

\[ \text{Singleton}(f) = \begin{cases} 1 & c_S(\overline{f}) = 1 \\ 0 & \text{otherwise} \end{cases} \]

\[ \text{Singleton}(f, e) = \begin{cases} 1 & c_S(\overline{f}, \overline{e}) = 1 \\ 0 & \text{otherwise} \end{cases} \]
Indicator features identify certain classes of rules

Features scored with $S$ and $L$:

$$\text{Singleton}(f) = \begin{cases} 
1 & c_S(\bar{f}) + c_L(\bar{f}) = 1 \\
0 & \text{otherwise}
\end{cases}$$

$$\text{Singleton}(f, e) = \begin{cases} 
1 & c_S(\bar{f}, \bar{e}) + c_L(\bar{f}, \bar{e}) = 1 \\
0 & \text{otherwise}
\end{cases}$$

$$\text{PostEditSupport}(f, e) = \begin{cases} 
1 & c_L(\bar{f}, \bar{e}) > 0 \\
0 & \text{otherwise}
\end{cases}$$
Choose feature weights that maximize objective function (BLEU score) on a development corpus

Minimum error rate training (MERT) (Och, 2003):

- Translate
- Optimize
Choose feature weights that maximize objective function (BLEU score) on a development corpus

Minimum error rate training (MERT) (Och, 2003):

Margin infused relaxed algorithm (MIRA) (Chiang 2012):
Post-Editing with Standard MT
Denkowski et al. (EACL 2014)

Static

Large LM  \[ X \rightarrow f/e \]  Weights

Grammar

Input Sentence  Decoder  Post-Editing
Post-Editing with Adaptive MT

Denkowski et al. (EACL 2014)

Static

Large Bitext

LM

Dynamic

PE Data

Weights

$w_1 \ldots w_n$

$X \rightarrow f/e$

TM

Input Sentence

Decoder

Post-Editing
How can we build systems without translators in the loop?
Overview

Online learning for statistical MT

- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments

- Kent State live post-editing

Automatic metrics for post-editing

- Meteor automatic metric
- Evaluation and optimization for post-editing

Conclusion
Simulated Post-Editing
Denkowski et al. (EACL 2014)

### Incremental training data

<table>
<thead>
<tr>
<th>Source</th>
<th>Target (Reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hola contestadora ...</td>
<td>Hello voicemail, my old ...</td>
</tr>
<tr>
<td>He llamado a servicio ...</td>
<td>I’ve called for tech ...</td>
</tr>
<tr>
<td>Ignoré la advertencia ...</td>
<td>I ignored my boss’ ...</td>
</tr>
<tr>
<td>Ahora anochece, y mi ...</td>
<td>Now it’s evening, and ...</td>
</tr>
<tr>
<td>Todavía sigo en espera ...</td>
<td>I’m still on hold. I’m ...</td>
</tr>
<tr>
<td>No creo que me hayas ...</td>
<td>I don’t think you ...</td>
</tr>
<tr>
<td>Ya he presionado cada ...</td>
<td>I punched every touch ...</td>
</tr>
</tbody>
</table>

Use **pre-generated references** in place of post-editing (Hardt and Elming, 2010)

Build, evaluate, and deploy adaptive systems using only **standard training data**
Simulated Post-Editing Experiments
Denkowski et al. (EACL 2014)

MT System (**cdec**)
- Hierarchical phrase-based model using **suffix arrays**
- Large **4-gram** language model
- **MIRA** optimization

Model Adaptation
- Update TM and weights **independently and in conjunction**

Training Data
- **WMT12** Spanish–English and **NIST 2012** Arabic–English

Evaluation Data
- WMT/NIST **news** (standard test sets)
- **TED talks** (totally blind out-of-domain test)
Simulated Post-Editing Experiments
Denkowski et al. (EACL 2014)

Spanish–English

Arabic–English

BLEU Score
Baseline
Grammar
MIRA
Both

NIST TED1 TED2
10
12
14
16
18
20
22
24
26
28

WMT TED1 TED2
26
28
30
32
34
36

Up to 1.7 BLEU improvement over static baseline
Simulated Post-Editing Experiments
Denkowski et al. (EACL 2014)

Spanish–English

Up to 1.7 BLEU improvement over static baseline

Arabic–English
Recent Work

How can we better leverage incremental data?
Translation Model Combination

Denkowski (AMTA 2014 Workshop on Interactive and Adaptive MT)

cdec (Dyer et al., 2010)

- Single translation model updated with new data
- Single feature set that changes over time (summation)

Moses (Koehn et al., 2007)

- Multiple translation models: background and post-editing
- Per-feature linear interpolation in context of full system

Recent additions to Moses toolkit

- Dynamic suffix array phrase tables (Germann, 2014)
- Fast MIRA implementation (Cherry and Foster, 2012)
- Multiple phrase tables with runtime weight updates (Denkowski, 2014)
Translation Model Combination
Denkowski (AMTA 2014 Workshop on Interactive and Adaptive MT)

Spanish–English

Arabic–English

Up to 4.9 BLEU improvement over static baseline
Up to 4.9 BLEU improvement over static baseline
Overview

Online learning for statistical MT

- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments

- Kent State live post-editing

Automatic metrics for post-editing

- Meteor automatic metric
- Evaluation and optimization for post-editing

Conclusion
Tools for Human Translators

Getting Started
Finding a location for your photo printer

- Place the photo printer on a flat, clean and dust-free surface, in a dry location, and out of direct sunlight.
- Allow at least 12 cm clearance from the back of the photo printer for the paper to travel.
- When connecting power or USB cables, keep the cables clear of the paper path to the front and rear of the photo printer.
- For proper ventilation, make sure the top and back of the photo printer are not blocked.
- Allow enough space on all sides of the photo printer to let you connect and disconnect cables, change the color cartridge, and add paper.

Connecting and turning on the power

Note:
Use only the AC power adapter included with your photo printer.
<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Translation</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Muchas gracias Chris. Y es en verdad un gran honor</td>
<td>Thank you so much, Chris. And it's truly a great honor</td>
<td>5 - Very Good</td>
</tr>
<tr>
<td>2</td>
<td>tener la oportunidad de venir a este escenario por segunda vez. Estoy extremadamente agradecido.</td>
<td>to have the opportunity to come to this stage twice. I'm extremely grateful.</td>
<td>1 - Gibberish</td>
</tr>
<tr>
<td>3</td>
<td>He quedado conmovido por esta conferencia, y deseo agradecer a todos ustedes</td>
<td>I have been moved by this conference, and I would like to thank all of you</td>
<td>Rate Translation</td>
</tr>
<tr>
<td>4</td>
<td>sus amables comentarios acerca de lo que tenía que decir la otra noche.</td>
<td></td>
<td>Rate Translation</td>
</tr>
<tr>
<td>5</td>
<td>Y digo eso sinceramente, en parte porque -- (Sollozos fingidos) -- ¡lo necesito! (Risas)</td>
<td></td>
<td>Rate Translation</td>
</tr>
<tr>
<td>6</td>
<td>¡Pónganse en mi posición!</td>
<td></td>
<td>Rate Translation</td>
</tr>
<tr>
<td></td>
<td>Source</td>
<td>Translation</td>
<td>Rating</td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>1</td>
<td>Muchas gracias Chris. Y es en verdad un gran honor</td>
<td>Thank you so much, Chris. And it's truly a great honor</td>
<td>5 - Very Good</td>
</tr>
<tr>
<td>2</td>
<td>tener la oportunidad de venir a este escenario por segunda vez. Estoy extremadamente agradecido.</td>
<td>to have the opportunity to come to this stage twice. I'm extremely grateful.</td>
<td>1 - Gibberish</td>
</tr>
<tr>
<td>3</td>
<td>He quedado conmovido por esta conferencia, y deseo agradecer a todos ustedes</td>
<td>I have been blown away by this conference, and I want to thank all of you for the many</td>
<td>5 - Very Good</td>
</tr>
<tr>
<td>4</td>
<td>sus amables comentarios acerca de lo que tenía que decir la otra noche.</td>
<td>Translating...</td>
<td>Rate Translation</td>
</tr>
<tr>
<td>5</td>
<td>Y digo eso sinceramente, en parte porque -- (Sollozos fingidos) -- ílo necesito! (Risas)</td>
<td></td>
<td>Rate Translation</td>
</tr>
<tr>
<td>6</td>
<td>¡Pónganse en mi posición!</td>
<td></td>
<td>Rate Translation</td>
</tr>
</tbody>
</table>
### TransCenter Post-Editing Interface

Denkowski and Lavie (AMTA 2012), Denkowski et al. (HaCat 2014)

<table>
<thead>
<tr>
<th>ID</th>
<th>MT</th>
<th>Post-Edited</th>
<th>Rating</th>
<th>Keypress</th>
<th>Mouseclick</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Thank you, Chris.</td>
<td>Thank you, Chris. And</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>22776</td>
</tr>
<tr>
<td>2</td>
<td>have the opportuni second time.</td>
<td>to have the opportuni second time. I am ext</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>26259</td>
</tr>
<tr>
<td>3</td>
<td>I have been move to thank all of you</td>
<td>I have been move by to thanks all of you</td>
<td>5</td>
<td>146</td>
<td>3</td>
<td>156690</td>
</tr>
<tr>
<td>4</td>
<td>for their kind com other night.</td>
<td>for your kind commer other night.</td>
<td>3</td>
<td>13</td>
<td>3</td>
<td>31397</td>
</tr>
<tr>
<td>5</td>
<td>And I say that sin fingidos) -- what</td>
<td>And I say that sincere -- I need it! (Laughter</td>
<td>2</td>
<td>39</td>
<td>3</td>
<td>48657</td>
</tr>
<tr>
<td>6</td>
<td>Put yourselves in</td>
<td>Put yourselves in my</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>21007</td>
</tr>
<tr>
<td>7</td>
<td>Volé on the plane</td>
<td>I flew on the vice pre</td>
<td>4</td>
<td>18</td>
<td>6</td>
<td>43915</td>
</tr>
<tr>
<td>8</td>
<td>Now I have my m plane!</td>
<td>Now I have take off m plane!</td>
<td>2</td>
<td>23</td>
<td>2</td>
<td>46021</td>
</tr>
<tr>
<td>9</td>
<td>(Laughter) (Applaus)</td>
<td>(Laughter) (Applause)</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1716</td>
</tr>
<tr>
<td>10</td>
<td>I will tell you a qu for me.</td>
<td>I will tell you a quick like for me.</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>14248</td>
</tr>
</tbody>
</table>
## TransCenter Post-Editing Interface

Denkowski and Lavie (AMTA 2012), Denkowski et al. (HaCat 2014)

<table>
<thead>
<tr>
<th>Time</th>
<th>Sentence 1 Edits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>The latest issue of Musharraf?</td>
</tr>
<tr>
<td>1345076800710</td>
<td>The latest issue of Musharraf?</td>
</tr>
<tr>
<td>1345076800861</td>
<td>The la of Musharraf?</td>
</tr>
<tr>
<td>1345076801036</td>
<td>The las of Musharraf?</td>
</tr>
<tr>
<td>1345076801212</td>
<td>The last of Musharraf?</td>
</tr>
<tr>
<td>1345076801341</td>
<td>The last of Musharraf?</td>
</tr>
<tr>
<td>1345076801508</td>
<td>The last a of Musharraf?</td>
</tr>
<tr>
<td>1345076801644</td>
<td>The last ac of Musharraf?</td>
</tr>
<tr>
<td>1345076801852</td>
<td>The last act of Musharraf?</td>
</tr>
<tr>
<td>1345076810117</td>
<td>The last act of Musharraf?</td>
</tr>
<tr>
<td>1345076811637</td>
<td>The of Musharraf?</td>
</tr>
<tr>
<td>1345076811796</td>
<td>The ofMusharraf?</td>
</tr>
<tr>
<td>1345076811973</td>
<td>The oMusharraf?</td>
</tr>
<tr>
<td>1345076814613</td>
<td>The Musharraf?</td>
</tr>
<tr>
<td>1345076815893</td>
<td>Musharraf?</td>
</tr>
<tr>
<td>1345076816100</td>
<td>Musharraf's?</td>
</tr>
<tr>
<td>1345076816269</td>
<td>Musharraf's?</td>
</tr>
<tr>
<td>1345076816535</td>
<td>Musharraf's last act?</td>
</tr>
<tr>
<td>Final</td>
<td>Musharraf's last act?</td>
</tr>
</tbody>
</table>
Overview

Online learning for statistical MT

- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments

- Kent State live post-editing

Automatic metrics for post-editing

- Meteor automatic metric
- Evaluation and optimization for post-editing

Conclusion
Experimental Setup

- Six translation studies students from Kent State University post-edited MT output

- Text: 4 excerpts from TED talks translated from Spanish into English (100 sentences total)

- Two excerpts translated by static system, two by adaptive system (shuffled by user)

- Record post-editing effort (HTER) and translator rating
Results

- Adaptive system **significantly outperforms** static baseline
- **Small improvement** in simulated scenario leads to **significant improvement** in production

<table>
<thead>
<tr>
<th></th>
<th>HTER ↓</th>
<th>Rating ↑</th>
<th>Sim PE BLEU ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>19.26</td>
<td>4.19</td>
<td>34.50</td>
</tr>
<tr>
<td>Adaptive</td>
<td>17.01</td>
<td>4.31</td>
<td><strong>34.95</strong></td>
</tr>
</tbody>
</table>
Overview

Online learning for statistical MT

- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments

- Kent State live post-editing

**Automatic metrics for post-editing**

- Meteor automatic metric
- Evaluation and optimization for post-editing

Conclusion
System Optimization

Parameter optimization (MIRA)

• Choose feature weights $W$ that maximizes objective on tuning set.

• Automatic metrics approximate human evaluation of MT output against reference translations.

Adequacy-based evaluation

• Good translations should be semantically similar to references.

• Several adequacy-driven research efforts:
  - ACL WMT (Callison-Burch et al., 2011)
  - NIST OpenMT (Przybocki et al., 2009)
Standard MT Evaluation

Standard BLEU metric based on $N$-gram precision ($P$) (Papineni et al., 2002)

- Matches spans of hypothesis $E'$ against reference $E$
- Surface forms only, depends on multiple references to capture translation variation (expensive)
- Jointly measures word choice and order

$$\text{BLEU} = BP \times \exp \left( \sum_{n=1}^{N} \frac{1}{N} \log P_n \right)$$

$$BP = \begin{cases} 1 & |E'| > |E| \\ e^{\frac{1-|E|}{|E'|}} & |E'| \leq |E| \end{cases}$$
Shortcomings of **BLEU** metric (Banerjee and Lavie 2005, Callison-Burch et al., 2007):

- Evaluating surface forms **misses correct translations**
- *N*-grams have no notion of **global coherence**
Shortcomings of **BLEU** metric (Banerjee and Lavie 2005, Callison-Burch et al., 2007):

- Evaluating surface forms *misses correct translations*
- *N*-grams have no notion of **global coherence**

**E:** The large home
Standard MT Evaluation

Shortcomings of BLEU metric (Banerjee and Lavie 2005, Callison-Burch et al., 2007):

- Evaluating surface forms misses correct translations
- $N$-grams have no notion of global coherence

$E$: The large home

$E': A$ big house  \hspace{1cm} \text{BLEU} = 0

$E'_1$: A big house  \hspace{1cm} \text{BLEU} = 0

$E'_2$: I am a dinosaur  \hspace{1cm} \text{BLEU} = 0
Post-Editing

Final translations must be human quality (editing required)

Good MT output should require less work for humans to edit

Human-targeted translation edit rate (HTER, Snover et al., 2006)

1. Human translators correct MT output
2. Automatically calculate number of edits using TER

$$\text{TER} = \frac{\# \text{ edits}}{|E|}$$

Edits: insertion, deletion, substitution, block shift

“Better” translations not always easier to post-edit
Translations scored by BLEU

E: The problem is that life of the lines is two to four years.
Translations scored by BLEU

\[ E: \text{ The problem is that life of the lines is two to four years.} \]

\[ E_1': \text{ The problem is that life is two lines, up to four years.} \]

\[ E_2': \text{ The problem is that the durability of lines is two or four years.} \]
Translations scored by BLEU

$E$: The problem is that life of the lines is two to four years.

$E'_1$: The problem is that life is two lines, up to four years.

$E'_2$: The problem is that the durability of lines is two or four years.
Translations scored by BLEU

\(E\):  The problem is that life of the lines is two to four years.

\(E'_1\):  The problem is that life of the lines, up to is two to four years.

\(E'_2\):  The problem is that the durability life of lines is two or to four years.
Overview

Online learning for statistical MT

- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments

- Kent State live post-editing

Automatic metrics for post-editing

- Meteor automatic metric
- Evaluation and optimization for post-editing

Conclusion
**Meteor**: alignment-based tunable evaluation metric

- Align hypothesis $E'$ to reference $E$

- Compute score based on alignment quality

Motivation: address shortcomings of BLEU

- Flexible matching to capture translation variation

- Measure word choice and order separately, combine with tunable scoring function

- Measure sentence coherence globally
The United States embassy know that dependable source.

The American embassy knows this from a reliable source.
The United States embassy knows that dependable source.

The American embassy knows this from a reliable source.

$E'$

$E$
The United States embassy knows that dependable
source.

The American embassy knows this from a reliable
source.
The United States embassy knows that dependable source.

The American embassy knows this from a reliable source.
The United States embassy know that dependable source.

The American embassy knows this from a reliable source.
The United States embassy knows that dependable source.

The American embassy knows this from a reliable source.

(P and R weighted by match type, content vs function words)
The United States embassy knows that dependable source.

The American embassy knows this from a reliable source.

Chunks = 2
Meteor Scoring
Denkowski and Lavie (2011)

$P$ and $R$ weighted by match type ($w_i, \ldots, w_n$) and content-function word weight ($\delta$)

$$F_\alpha = \frac{P \times R}{\alpha \times P + (1 - \alpha) \times R}$$  \hspace{1cm}  \text{Frag} = \frac{\text{Chunks}}{\text{AvgMatches}}$$

$$\text{Meteor} = \left(1 - \gamma \times \text{Frag}^\beta\right) \times F_\alpha$$

Tunable parameters:

- $W = \langle w_i, \ldots, w_n \rangle$: weights for flexible match types
- $\alpha$: balance between precision and recall
- $\beta, \gamma$: weight and severity of fragmentation
- $\delta$: relative contribution of content versus function words
Casting Meteor’s scoring features as post-editing measures:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>Incorrect content (deletion)</td>
</tr>
<tr>
<td>Recall</td>
<td>Missing content (insertion)</td>
</tr>
<tr>
<td>Fragmentation</td>
<td>Incorrectly ordered content (reordering)</td>
</tr>
<tr>
<td>Match types</td>
<td>Partially correct content (minor edits)</td>
</tr>
<tr>
<td>Content vs Function</td>
<td>Content vs grammaticality edits</td>
</tr>
</tbody>
</table>

Advantage over edit distance: error types identified separately and combined with a parameterized scoring function.
Overview

Online learning for statistical MT
- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments
- Kent State live post-editing

Automatic metrics for post-editing
- Meteor automatic metric
  - Evaluation and optimization for post-editing

Conclusion
Metrics Targeting Post-Editing
Denkowski (AMTA 2014 Workshop on Interactive and Adaptive MT)

Startup

- Deploy system tuned with simulated post-editing and BLEU
- Collect enough data for a post-editing dev set

Retuning (second stage booster rocket)

- Tune Meteor to fit post-editing effort (keystroke, very close to rating)
- Tune system to new Meteor on new dev set
- Continue to adapt to Meteor in production
Second Field Test
Denkowski (AMTA 2014 Workshop on Interactive and Adaptive MT)

Results

- Repeat post-editing experiments with second set of students and TED talks

- Compare BLEU and Meteor-tuned adaptive systems (both optimized on TED talk data)

- Adapting to Meteor lowers BLEU but yields significant improvement in live post-editing

- Feasible in production: significant data and editing records

<table>
<thead>
<tr>
<th></th>
<th>HTER ↓</th>
<th>Rating ↑</th>
<th>Sim PE BLEU ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adapt BLEU</td>
<td>20.1</td>
<td>4.16</td>
<td>27.3</td>
</tr>
<tr>
<td>Adapt Meteor</td>
<td>18.9</td>
<td>4.24</td>
<td>26.6</td>
</tr>
</tbody>
</table>
Overview

Online learning for statistical MT

- Translation model review
- Real time model adaptation
- Simulated post-editing

Post-editing software and experiments

- Kent State live post-editing

Automatic metrics for post-editing

- Meteor automatic metric
- Evaluation and optimization for post-editing

Conclusion
Conclusion

Real time adaptive MT systems

- **Immediately incorporate** post-editing data into translation models
- **Run an** online optimizer **that continuously updates feature weights during decoding**
- **Simulate** post-editing to train on normal system building data
- **See best results when combining techniques**: up to 4.9 BLEU
Live post-editing experiments

- TransCenter interface that simplifies and records post-editing tasks

- Live experiments that show a reduction in human labor when working with adaptive systems
Conclusion

The United States embassy knows that dependable source.

The American embassy knows this from a reliable source.

Automatic metrics for post-editing

- Meteor MT evaluation metric capable of fitting various measures of editing effort
- Live experiments that show a further gains in translator productivity when systems adapt to Meteor
Building adaptive MT systems

- **cdec Realtime**: adaptive MT systems with cdec
- **RTA**: Realtime adaptive MT framework using Moses

Live post-editing

- **TransCenter**: post-editing data collection interface
- **All Kent State** post-editing data

Targeted automatic metrics

- **Meteor**: tunable MT evaluation metric
Adaptive MT Systems for Post-Editing Tasks
Machine Translation for Human Translators

Michael Denkowski, Alon Lavie, Chris Dyer, Jaime Carbonell, Gregory Shreve*, Isabel Lacruz*

Carnegie Mellon University, *Kent State University

May 13, 2015