Recent Trends in Computer Aided Translation

Philipp Koehn

16 September 2016
Overview

• A practical introduction: the CASMACAT workbench

• Postediting

• Types of assistance

• Logging, eye tracking and user studies
CASMACAT Project 2011-2014

• Cognitive studies of translators leading to insights into interface design
  → better understanding of translator needs

• Workbench with novel types of assistance to human translators
  – interactive translation prediction
  – interactive editing and reviewing
  – adaptive translation models
  → better tools for translators

• Demonstration of effectiveness in field tests with professional translators
  → increased translator productivity
• Source on left, translation on right
• Context above and below
Confidence Measures

- Sentence-level confidence measures
  → estimate usefulness of machine translation output

- Word-level confidence measures
  → point posteditor to words that need to be changed
Incremental Updating
Incremental Updating

Machine Translation

Postediting
Incremental Updating
Pour la science, cela sert à vérifier la validité du Modèle standard (MS), et cela permet aussi aux physiciens de scruter tout écart entre les observations et les prédictions du MS.

 Ils sont d'ailleurs plusieurs à souhaiter ardemment qu'on en trouve, car la moindre différence pourrait ouvrir une porte sur une "nouvelle physique" et boucher certains trous du Modèle.
Pour mieux redistribuer ses cartes, Moucharraf a envoyé l'armée pakistanaise dans les zones ethniques qui longent l'Afghanistan, pour la première fois depuis l'indépendance du Pakistan.

In furtherance of his re-alignment, Musharraf sent the Pakistani army into the tribal areas bordering Afghanistan for the first time since Pakistan's independence.
Word Alignment

Ils sont d'ailleurs plusieurs à souhaiter ardiment qu'on en trouve, car la moindre différence pourrait ouvrir une porte sur une "nouvelle physique" et boucher certains trous du Modèle.

There are elsewhere several who wish fervently that,, because

- With interactive translation prediction
- Shade off translated words, highlight next word to translate
Translation Option Array

- Visual aid: non-intrusive provision of cues to the translator
- Clickable: click on target phrase → added to edit area
- Automatic orientation
  - most relevant is next word to be translated
  - automatic centering on next word
**Bilingual Concordancer**

**abandonner**

**abandon**

- inces des Etats-Unis à abandonner Musharraf -- et les con..
- juridique, il a décidé d'abandonner la constitutionnalité, d'
- implement menacé d’abandonner ses accords commerci
- simply threatened to abandon or never to conclude t

**give up**

- arait donc contraint d’abandonner le droit de créer son p
- n’ était pas disposé à abandonner ses fonctions militaire
- erraf was not ready to give up his military post, but a

**to**

- t ne veulent donc pas abandonner leurs prérogatives dar
- olicy and do not want to delegate this prerogat

**to abandon**

- es tout en refusant d’abandonner son arsenal nucléaire
- drawal while refusing to abandon its nuclear weapons a
However, the European Central Bank (ECB) asked about it in a report on virtual currencies published in October.

Paraphrases for "However"

- on the other hand
- nevertheless
How do we Know it Works?

- **Intrinsic Measures**
  - word level confidence: user does not change words generated with certainty
  - interactive prediction: user accepts suggestions

- **User Studies**
  - professional translators faster with post-editing
  - ... but like interactive translation prediction better

- **Cognitive studies with eye tracking**
  - where is the translator looking at?
  - what causes the translator to be slow?
Logging and Eye Tracking

Pre-loading MT suggestion

Reading TT segment

Reading ST segment

Post-Editing activities (ins, del)
Home Edition

- Running **CASMACAT** on your desktop or laptop

- **Installation**
  - Installation software to run virtual machines (e.g., Virtualbox)
  - installation of Linux distribution (e.g., Ubuntu)
  - installation script sets up all the required software and dependencies
Administration through Web Browser

Administration

Translate
- Translate new document
- List documents

Engines
- Manage engines
- Upload engine
- Build new prototype

Settings
- Reset CAT and MT server
- CAT Settings
- Update Software

Deployed: fr-en-upload-1
Memory: 1.2 GB used, 6.6 GB free
Disk: 12.9 GB used, 10.2 GB free
Uptime: 22:24
Load: 0.01, 0.05, 0.08
Monday, 06 October 2014, 21:22:41
Training MT Engines

- Train MT engine on own or public data
### Managing MT Engines

#### Manage Engines

#### English-French

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Size</th>
<th>Build date</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>NC+TED</td>
<td>2.3G</td>
<td>27 Mar 14</td>
<td><strong>deploy</strong> <strong>delete</strong> <strong>download</strong></td>
</tr>
</tbody>
</table>

**Prototypes** *(Inspect Details in Prototype Factory)*

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Status</th>
<th>Build date</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>NC+TED</td>
<td>done</td>
<td>Fri 20:34</td>
<td>delete</td>
</tr>
<tr>
<td>1</td>
<td>NC</td>
<td>done</td>
<td>Fri 20:34</td>
<td>create engine delete</td>
</tr>
</tbody>
</table>

#### English-Spanish

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Size</th>
<th>Build date</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>NC+TED</td>
<td>2.3G</td>
<td>27 Mar 14</td>
<td><strong>deploy</strong> <strong>delete</strong> <strong>download</strong></td>
</tr>
</tbody>
</table>

**Prototypes** *(Inspect Details in Prototype Factory)*

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Status</th>
<th>Build date</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>NC+TED+EP</td>
<td>stopped</td>
<td>Fri 20:34</td>
<td>resume delete</td>
</tr>
<tr>
<td>2</td>
<td>NC+TED</td>
<td>done</td>
<td>Fri 20:34</td>
<td>delete</td>
</tr>
<tr>
<td>1</td>
<td>NC</td>
<td>done</td>
<td>Fri 20:34</td>
<td>create engine delete</td>
</tr>
</tbody>
</table>
part II

cat methods
post-editing
Productivity Improvements

(source: Autodesk)
## MT Quality and Productivity

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>Training Sentences</th>
<th>Training Words (English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT1</td>
<td>30.37</td>
<td>14,700k</td>
<td>385m</td>
</tr>
<tr>
<td>MT2</td>
<td>30.08</td>
<td>7,350k</td>
<td>192m</td>
</tr>
<tr>
<td>MT3</td>
<td>29.60</td>
<td>3,675k</td>
<td>96m</td>
</tr>
<tr>
<td>MT4</td>
<td>29.16</td>
<td>1,837k</td>
<td>48m</td>
</tr>
<tr>
<td>MT5</td>
<td>28.61</td>
<td>918k</td>
<td>24m</td>
</tr>
<tr>
<td>MT6</td>
<td>27.89</td>
<td>459k</td>
<td>12m</td>
</tr>
<tr>
<td>MT7</td>
<td>26.93</td>
<td>230k</td>
<td>6.0m</td>
</tr>
<tr>
<td>MT8</td>
<td>26.14</td>
<td>115k</td>
<td>3.0m</td>
</tr>
<tr>
<td>MT9</td>
<td>24.85</td>
<td>57k</td>
<td>1.5m</td>
</tr>
</tbody>
</table>

- Same type of system (Spanish–English, phrase-based, Moses)
- Trained on varying amounts of data [Sanchez-Torron and Koehn, AMTA 2016]
## MT Quality and Productivity

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>Training Sentences</th>
<th>Training Words (English)</th>
<th>Post-Editing Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT1</td>
<td>30.37</td>
<td>14,700k</td>
<td>385m</td>
<td>4.06 sec/word</td>
</tr>
<tr>
<td>MT2</td>
<td>30.08</td>
<td>7,350k</td>
<td>192m</td>
<td>4.38 sec/word</td>
</tr>
<tr>
<td>MT3</td>
<td>29.60</td>
<td>3,675k</td>
<td>96m</td>
<td>4.23 sec/word</td>
</tr>
<tr>
<td>MT4</td>
<td>29.16</td>
<td>1,837k</td>
<td>48m</td>
<td>4.54 sec/word</td>
</tr>
<tr>
<td>MT5</td>
<td>28.61</td>
<td>918k</td>
<td>24m</td>
<td>4.35 sec/word</td>
</tr>
<tr>
<td>MT6</td>
<td>27.89</td>
<td>459k</td>
<td>12m</td>
<td>4.36 sec/word</td>
</tr>
<tr>
<td>MT7</td>
<td>26.93</td>
<td>230k</td>
<td>6.0m</td>
<td>4.66 sec/word</td>
</tr>
<tr>
<td>MT8</td>
<td>26.14</td>
<td>115k</td>
<td>3.0m</td>
<td>4.94 sec/word</td>
</tr>
<tr>
<td>MT9</td>
<td>24.85</td>
<td>57k</td>
<td>1.5m</td>
<td>5.03 sec/word</td>
</tr>
</tbody>
</table>

- User study with professional translators
- Correlation between BLEU and post-editing speed?
MT Quality and Productivity

BLEU against PE speed and regression line with 95% confidence bounds
+1 BLEU ↔ decrease in PE time of ~0.16 sec/word, or 3-4% speed-up
MT Quality and PE Quality

better MT $\leftrightarrow$ fewer post-editing errors
## Translator Variability

<table>
<thead>
<tr>
<th></th>
<th>HTER</th>
<th>Edit Rate</th>
<th>PE speed (spw)</th>
<th>MQM Score</th>
<th>Fail</th>
<th>Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR1</td>
<td>44.79</td>
<td>2.29</td>
<td>4.57</td>
<td>98.65</td>
<td>10</td>
<td>124</td>
</tr>
<tr>
<td>TR2</td>
<td>42.76</td>
<td>3.33</td>
<td>4.14</td>
<td>97.13</td>
<td>23</td>
<td>102</td>
</tr>
<tr>
<td>TR3</td>
<td>34.18</td>
<td>2.05</td>
<td>3.25</td>
<td>96.50</td>
<td>26</td>
<td>106</td>
</tr>
<tr>
<td>TR4</td>
<td>49.90</td>
<td>3.52</td>
<td>2.98</td>
<td>98.10</td>
<td>17</td>
<td>120</td>
</tr>
<tr>
<td>TR5</td>
<td>54.28</td>
<td>4.72</td>
<td>4.68</td>
<td>97.45</td>
<td>17</td>
<td>119</td>
</tr>
<tr>
<td>TR6</td>
<td>37.14</td>
<td>2.78</td>
<td><strong>2.86</strong></td>
<td>97.43</td>
<td>24</td>
<td>113</td>
</tr>
<tr>
<td>TR7</td>
<td>39.18</td>
<td>2.23</td>
<td><strong>6.36</strong></td>
<td>97.92</td>
<td>18</td>
<td>112</td>
</tr>
<tr>
<td>TR8</td>
<td>50.77</td>
<td>7.63</td>
<td>6.29</td>
<td>97.20</td>
<td>19</td>
<td>117</td>
</tr>
<tr>
<td>TR9</td>
<td>39.21</td>
<td>2.81</td>
<td>5.45</td>
<td>96.48</td>
<td>22</td>
<td>113</td>
</tr>
</tbody>
</table>

- Higher variability between translators than between MT systems
confidence measures

(“quality estimation”)
• Machine translation engine indicates where it is likely wrong

• Different Levels of granularity
  – document-level (SDL’s “TrustScore”)
  – sentence-level
  – word-level
Sentence-Level Confidence

- Translators are used to “Fuzzy Match Score”
  - used in translation memory systems
  - roughly: ratio of words that are the same between input and TM source
  - if less than 70%, then not useful for post-editing

- We would like to have a similar score for machine translation

- Even better
  - estimation of post-editing time
  - estimation of from-scratch translation time
  → can also be used for pricing

- Very active research area
Quality Estimation Shared Task

- Shared task organized at WMT since 2012

- Given
  - source sentence
  - machine translation

- Predict
  - human judgement of usefulness for post-editing (2012, 2014)
  - HTER score on post-edited sentences (2013–2016)
  - post-editing time (2013, 2014)

- Also task for word-level quality estimation (2014–2016) and document-level quality estimation (2015)
• Open source tool for quality estimation

• Source sentence features
  – number of tokens
  – language model (LM) probability
  – 1–3-grams observed in training corpus
  – average number of translations per word

• Similar target sentence features

• Alignment features
  – difference in number of tokens and characters
  – ratio of numbers, punctuation, nouns, verbs, named entities
  – syntactic similarity (POS tags, constituents, dependency relationships)

• Scores and properties of the machine translation derivation

• Uses Python’s scikit-learn implementation of SVM regression
WMT 2016: Best System

- Yandex School of Data Analysis (Kozlova et al., 2016)

- QuEst approach with additional features
  - syntactically motivated features
  - language model and statistics on web-scale corpus
  - pseudo-references and back-translations
  - other miscellaneous features

- Performance
  - mean average HTER difference 13.53
  - ranking correlation 0.525
word level confidence
• Highlight words less likely to be correct
Methods

• Simple methods quite effective
  – IBM Model 1 scores
  – posterior probability of the MT model

• Machine learning approach
  – similar features as for sentence-level quality estimation
• Machine translation output

Quick brown fox jumps on the dog lazy.

• Post-editing

The quick brown fox jumps over the lazy dog.

• Annotation

Fast brown fox jumps on the dog lazy.
bad good good good bad good good good good

• Problems: dropped words? reordering?
Quality Requirements

• Evaluated in user study

• Feedback
  – could be useful feature
  – but accuracy not high enough

• To be truly useful, accuracy has to be very high

• Current methods cannot deliver this
WMT 2016: Best System

• Unbabel (Martins et al., 2016)

• Viewed as tagging task

• Features: black box and language model features

• Method: Combination of
  – feature-rich linear HMM model
  – deep neural networks
    (feed-forward, bi-directionally recurrent, convolutional)

• Performance
  – F-score for detecting good words: 88.45
  – F-score for detecting bad words: 55.99
interactive translation prediction
Input Sentence

Er hat seit Monaten geplant, im Oktober einen Vortrag in Miami zu halten.

Professional Translator
Input Sentence

Er hat seit Monaten geplant, im Oktober einen Vortrag in Miami zu halten.

Professional Translator

| He |
Input Sentence
Er hat seit Monaten geplant, im Oktober einen Vortrag in Miami zu halten.

Professional Translator
He | has
**Input Sentence**

Er hat seit Monaten geplant, im Oktober einen Vortrag in Miami zu halten.

**Professional Translator**

He has | for months
Input Sentence

Er hat seit Monaten geplant, im Oktober einen Vortrag in Miami zu halten.

Professional Translator

He planned |
Input Sentence
Er hat seit Monaten geplant, im Oktober einen Vortrag in Miami zu halten.

Professional Translator
He planned | for months
Visualization

• Show \( n \) next words

Olvidarlo. Es demasiado | arriesgado. Estoy haciendo

• Show rest of sentence
Spence Green’s Lilt System

• Show alternate translation predictions

• Show alternate translations predictions with probabilities
Search for best translation creates a graph of possible translations
One path in the graph is the best (according to the model)

This path is suggested to the user
The user may enter a different translation for the first words

We have to find it in the graph
We can predict the optimal completion (according to the model)
- Average response time based on length of the prefix and number of edits
- Main bottleneck is the string edit distance between prefix and path.
Word Completion

• Complete word once few letters are typed

• Example: predict college over university?

• User types the letter $u \rightarrow$ change prediction

• ”Desperate” word completion: find any word that matches
Redecoding

• Translate the sentence again, enforce matching the prefix

• Recent work on this: Wuebker et al. [ACL 2016]

Models and Inference for Prefix-Constrained Machine Translation

Joern Wuebker, Spence Green, John DeNero, Saša Hasan
Lilt, Inc.
first_name@lilt.com

Minh-Thang Luong
Stanford University
lmthang@stanford.edu
Prefix-Matching Decoding

• Prefix-matching phase
  – only allow translation options that match prefix
  – prune based on target words matched

• Ensure that prefix can be created by system
  – add synthetic translation options from word aligned prefix
    (but with low probability)
  – no reordering limit

• After prefix is match, regular beam search

• Fast enough?
  \(\Rightarrow\) Wuebker et al. [ACL 2016] report 51-89ms per sentence
Tuning

• Optimize to produce better predictions

• Focus on next few words, not full sentence

• Tuning metric
  – prefix BLEU (ignoring prefix to measure score)
  – word prediction accuracy
  – length of correctly predicted suffix sequence

• Generate diverse n-best list to ensure learnability

• Wuebker et al. [ACL 2016] report significant gains
Neural Interactive Translation Prediction

- Recent success of neural machine translation
- For instance, attention model
Neural MT: Sequential Prediction

- The model produces words in sequence

\[ p(\text{output}_t|\{\text{output}_1, \cdots, \text{output}_{t-1}\}, \text{input}) = g(\hat{\text{output}}_{t-1}, \text{context}_t, \text{hidden}_t) \]

- Translation prediction: feed in user prefix
Das Unternehmen sagte, dass es in diesem Monat mit Bewerbungsgesprächen beginnen wird und die Mitarbeiterzahl von Oktober bis Dezember steigt.
- Better prediction accuracy, even when systems have same BLEU score (state-of-the-art German-English systems, compared to search graph matching)

<table>
<thead>
<tr>
<th>System</th>
<th>Configuration</th>
<th>BLEU</th>
<th>Word Prediction Accuracy</th>
<th>Letter Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural</td>
<td>no beam search</td>
<td>34.5</td>
<td>61.6%</td>
<td>86.8%</td>
</tr>
<tr>
<td></td>
<td>beam size 12</td>
<td>36.2</td>
<td>63.6%</td>
<td>87.4%</td>
</tr>
<tr>
<td>Phrase-based</td>
<td>-</td>
<td>34.5</td>
<td>43.3%</td>
<td>72.8%</td>
</tr>
</tbody>
</table>
Recovery from Failure

- Ratio of words correct after first failure

<table>
<thead>
<tr>
<th>System</th>
<th>Configuration</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural</td>
<td>no beam search</td>
<td>55.9%</td>
<td>61.8%</td>
<td>61.3%</td>
<td>62.2%</td>
<td>61.1%</td>
</tr>
<tr>
<td></td>
<td>beam size 12</td>
<td>58.0%</td>
<td>62.9%</td>
<td>62.8%</td>
<td>64.0%</td>
<td>61.5%</td>
</tr>
<tr>
<td>Phrase-based</td>
<td>-</td>
<td>28.6%</td>
<td>45.5%</td>
<td>46.9%</td>
<td>47.4%</td>
<td>48.4%</td>
</tr>
</tbody>
</table>

- Depending on probability of user word (neural, no beam)
Patching Translations

- Decoding speeds
  - translation speed with CPU: 100 ms/word
  - translation speed with GPU: 7ms/word

- To stay within 100ms speed limit
  - predict only a few words ahead (say, 5, in $5 \times 7\text{ms} = 35\text{ms}$)
  - patch new partial prediction with old full sentence prediction
  - uses KL divergence to find best patch point in $\pm 2$ word window

- May compute new full sentence prediction in background, return as update

- Only doing quick response reduces word prediction accuracy 61.6% $\rightarrow$ 56.4%
translation options
Translation Option Array

- Visual aid: non-intrusive provision of cues to the translator
- Trigger passive vocabulary
How to Rank

- Basic idea: best options on top

- Problem: how to rank word translation vs. phrase translations?

- Method: utilize future cost estimates

- Translation score
  - sum of translation model costs
  - language model estimate
  - outside future cost estimate

\[
\begin{align*}
\text{tm:} & \text{-0.56, lm:} \text{-2.81} \\
\text{d:} & \text{-0.74. all:} \text{-4.11} \\
\text{-9.3 +} & \\
\text{-4.11} & \text{=} \\
\text{-13.41}
\end{align*}
\]
Improving Rankings

- Removal of duplicates and near duplicates

<table>
<thead>
<tr>
<th>bad</th>
<th>good</th>
</tr>
</thead>
<tbody>
<tr>
<td>erupted</td>
<td>climbing</td>
</tr>
<tr>
<td>ausbrach</td>
<td>Klettern</td>
</tr>
<tr>
<td>ausbrach,</td>
<td>Bergsteigen</td>
</tr>
<tr>
<td>platzte</td>
<td>Aufstieg</td>
</tr>
<tr>
<td>Ausbruch</td>
<td>abhalten,</td>
</tr>
<tr>
<td>ausgebrogen</td>
<td>Erklimmen</td>
</tr>
<tr>
<td>ausgebrogen ist</td>
<td>beim Besteigen</td>
</tr>
</tbody>
</table>

- Ranking by likelihood to be used in the translation
  → can this be learned from user feedback?
Enabling Monolingual Translators

- Monolingual translator
  - wants to understand a foreign document
  - has no knowledge of foreign language
  - uses a machine translation system

- Questions
  - Is current MT output sufficient for understanding?
  - What else could be provided by a MT system?
• MT system output:

    The study also found that one of the genes in the improvement in people with prostate cancer risk, it also reduces the risk of suffering from diabetes.

• What does this mean?

• Monolingual translator:

    The research also found that one of the genes increased people’s risk of prostate cancer, but at the same time lowered people’s risk of diabetes.

• Document context helps
Example: Arabic

<table>
<thead>
<tr>
<th>Arabic</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>the the us house of representatives</td>
<td>adopted thursday legally calls for the withdrawal of combat troops from iraq in april</td>
</tr>
<tr>
<td>the us house of representatives</td>
<td>adopted the thu the legally demands withdrawal of troops from iraq in the latest april</td>
</tr>
<tr>
<td>it was the us house of representatives</td>
<td>was adopted thursday the the law demands withdrawal of troops from iraq in april</td>
</tr>
<tr>
<td>he</td>
<td>was adopted by the us house a legally calls for withdrawal of forces the fighter from the first on april</td>
</tr>
<tr>
<td>earlier</td>
<td>a</td>
</tr>
<tr>
<td>it was the us house</td>
<td>was adopted thursday the the legal demands withdrawal of forces the fighter from the first on april</td>
</tr>
<tr>
<td>earlier, the</td>
<td>adopted, the thu, the legal calls for withdrawal of forces the fighter from the first on april</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Arabic</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008, defy</td>
<td>once new president george w. bush, who opposes no date has been set for</td>
</tr>
<tr>
<td>the 2008 defiant</td>
<td>once again president george bush who opposes no date has been set for</td>
</tr>
<tr>
<td>2008, challenging</td>
<td>again the new president george bush who opposes no date has been set for</td>
</tr>
<tr>
<td>a defiant</td>
<td>the first president george bush who opposes no date has been set for</td>
</tr>
<tr>
<td>In defiance of</td>
<td>once again, a defiant the first president george bush who opposes no date has been set for</td>
</tr>
<tr>
<td>In defiance of the</td>
<td>once again the new president george bush who opposes no date has been set for</td>
</tr>
<tr>
<td>In 2008</td>
<td>challenging the time the new president george w. bush 's who opposes no date has been set for</td>
</tr>
<tr>
<td>challenging the time, defy</td>
<td>once again, the new president george w. bush 's who opposes no date has been set for</td>
</tr>
</tbody>
</table>

up to 10 translations for each word / phrase
### Example: Arabic

<table>
<thead>
<tr>
<th>بالسحب</th>
<th>القوات</th>
<th>المقاتلة</th>
<th>التابعة</th>
<th>العراق</th>
</tr>
</thead>
<tbody>
<tr>
<td>withdrawal of</td>
<td>combat troops</td>
<td>the fighting forces</td>
<td>us</td>
<td>Iraq</td>
</tr>
<tr>
<td>بالسحب</td>
<td>القوات</td>
<td>المقاتلة</td>
<td>التابعة</td>
<td>العراق</td>
</tr>
<tr>
<td>withdrawal of</td>
<td>combat troops</td>
<td>the fighting forces</td>
<td>us</td>
<td>Iraq</td>
</tr>
<tr>
<td>بالسحب</td>
<td>القوات</td>
<td>المقاتلة</td>
<td>التابعة</td>
<td>عراق</td>
</tr>
<tr>
<td>withdrawal of</td>
<td>troops</td>
<td>fighter</td>
<td>the us</td>
<td>Iraq</td>
</tr>
<tr>
<td>بالسحب</td>
<td>القوات</td>
<td>المقاتلة</td>
<td>التابعة</td>
<td>عراق</td>
</tr>
<tr>
<td>withdrawal of</td>
<td>combat forces</td>
<td>the fighter</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Monolingual Translation with Options

No big difference — once significantly better
Monolingual Translation Triage

• Study on Russian–English (Schwartz, 2014)

• Allow monolingual translators to assess their translation
  - confident $\rightarrow$ accept the translation
  - verify $\rightarrow$ proofread by bilingual
  - partially unsure $\rightarrow$ part of translation handled by bilingual
  - completely unsure $\rightarrow$ handled by bilingual

• Monolingual translator highly effective in triage
Monolingual Translation: Conclusions

• Main findings
  – monolingual translators may be as good as bilinguals
  – widely different performance by translator / story
  – named entity translation critically important

• Various human factors important
  – domain knowledge
  – language skills
  – effort
logging and eye tracking
Logging functions

- Different types of events are saved in the logging.
  - configuration and statistics
  - start and stop session
  - segment opened and closed
  - text, key strokes, and mouse events
  - scroll and resize
  - search and replace
  - suggestions loaded and suggestion chosen
  - interactive translation prediction
  - gaze and fixation from eye tracker
Logging functions

• In every event we save:
  – Type
  – In which element was produced
  – Time

• Special attributes are kept for some types of events
  – Diff of a text change
  – Current cursor position
  – Character looked at
  – Clicked UI element
  – Selected text

⇒ Full replay of user session is possible
Keystroke Log

Input:  Au premier semestre, l’avionneur a livré 97 avions.
Output: The manufacturer has delivered 97 planes during the first half.

(37.5 sec, 3.4 sec/word)

black: keystroke, purple: deletion, grey: cursor move
height: length of sentence
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Src.</td>
<td>Sans se démonter, il s’est montré concis et précis.</td>
<td>MT</td>
</tr>
<tr>
<td>1/3</td>
<td>Without fail, he has been concise and accurate.</td>
<td>(Prediction+Options, L2a)</td>
</tr>
<tr>
<td>4/0</td>
<td>Without getting flustered, he showed himself to be concise and precise. (Unassisted, L2b)</td>
<td></td>
</tr>
<tr>
<td>4/0</td>
<td>Without falling apart, he has shown himself to be concise and accurate. (Postedit, L2c)</td>
<td></td>
</tr>
<tr>
<td>1/3</td>
<td>Unswayable, he has shown himself to be concise and to the point. (Options, L2d)</td>
<td></td>
</tr>
<tr>
<td>0/4</td>
<td>Without showing off, he showed himself to be concise and precise. (Prediction, L2e)</td>
<td></td>
</tr>
<tr>
<td>1/3</td>
<td>Without dismantling himself, he presented himself consistent and precise. (Prediction+Options, L1a)</td>
<td></td>
</tr>
<tr>
<td>2/2</td>
<td>He showed himself concise and precise. (Unassisted, L1b)</td>
<td></td>
</tr>
<tr>
<td>3/1</td>
<td>Nothing daunted, he has been concise and accurate. (Postedit, L1c)</td>
<td></td>
</tr>
<tr>
<td>3/1</td>
<td>Without losing face, he remained focused and specific. (Options, L1d)</td>
<td></td>
</tr>
<tr>
<td>3/1</td>
<td>Without becoming flustered, he showed himself concise and precise. (Prediction, L1e)</td>
<td></td>
</tr>
</tbody>
</table>

**Example of Quality Judgments**
## Main Measure: Productivity

<table>
<thead>
<tr>
<th>Assistance</th>
<th>Speed</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unassisted</td>
<td>4.4s/word</td>
<td>47% correct</td>
</tr>
<tr>
<td>Postedit</td>
<td>2.7s (-1.7s)</td>
<td>55% (+8%)</td>
</tr>
<tr>
<td>Options</td>
<td>3.7s (-0.7s)</td>
<td>51% (+4%)</td>
</tr>
<tr>
<td>Prediction</td>
<td>3.2s (-1.2s)</td>
<td>54% (+7%)</td>
</tr>
<tr>
<td>Prediction+Options</td>
<td>3.3s (-1.1s)</td>
<td>53% (+6%)</td>
</tr>
</tbody>
</table>
## Faster and Better, Mostly

<table>
<thead>
<tr>
<th>User</th>
<th>Unassisted</th>
<th>Postedit</th>
<th>Options</th>
<th>Prediction</th>
<th>Prediction+Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1a</td>
<td>3.3 sec/word 23% correct</td>
<td>1.2s -2.2s 39% +16%</td>
<td>2.3s -1.0s 45% +22%</td>
<td>1.1s -2.2s 30% +7%</td>
<td>2.4s -0.9s 44% +21%</td>
</tr>
<tr>
<td>L1b</td>
<td>7.7 sec/word 35% correct</td>
<td>4.5s -3.2s 48% +13%</td>
<td>4.5s -3.3s 55% +20%</td>
<td>2.7s -5.1s 61% +26%</td>
<td>4.8s -3.0s 41% +6%</td>
</tr>
<tr>
<td>L1c</td>
<td>3.9 sec/word 50% correct</td>
<td>1.9s -2.0s 61% +11%</td>
<td>3.8s -0.1s 54% +4%</td>
<td>3.1s -0.8s 64% +14%</td>
<td>2.5s -1.4s 61% +11%</td>
</tr>
<tr>
<td>L1d</td>
<td>2.8 sec/word 38% correct</td>
<td>2.0s -0.7s 46% +8%</td>
<td>2.9s (+0.1s) 59% (+21%)</td>
<td>2.4s (-0.4s) 37% (-1%)</td>
<td>1.8s -1.0s 45% +7%</td>
</tr>
<tr>
<td>L1e</td>
<td>5.2 sec/word 58% correct</td>
<td>3.9s -1.3s 64% +6%</td>
<td>4.9s (-0.2s) 56% (-2%)</td>
<td>3.5s -1.7s 62% +4%</td>
<td>4.6s (-0.5s) 56% (-2%)</td>
</tr>
<tr>
<td>L2a</td>
<td>5.7 sec/word 16% correct</td>
<td>1.8s -3.9s 50% +34%</td>
<td>2.5s -3.2s 34% +18%</td>
<td>2.7s -3.0s 40% +24%</td>
<td>2.8s -2.9s 50% +34%</td>
</tr>
<tr>
<td>L2b</td>
<td>3.2 sec/word 64% correct</td>
<td>2.8s (-0.4s) 56% (-8%)</td>
<td>3.5s +0.3s 60% -4%</td>
<td>6.0s +2.8s 61% -3%</td>
<td>4.6s +1.4s 57% -7%</td>
</tr>
<tr>
<td>L2c</td>
<td>5.8 sec/word 52% correct</td>
<td>2.9s -3.0s 53% +1%</td>
<td>4.6s (-1.2s) 37% (-15%)</td>
<td>4.1s -1.7s 59% +7%</td>
<td>2.7s -3.1s 53% +1%</td>
</tr>
<tr>
<td>L2d</td>
<td>3.4 sec/word 49% correct</td>
<td>3.1s (-0.3s) 49% (+0%)</td>
<td>4.3s (+0.9s) 51% (+2%)</td>
<td>3.8s (+0.4s) 53% (+4%)</td>
<td>3.7s (+0.3s) 58% (+9%)</td>
</tr>
<tr>
<td>L2e</td>
<td>2.8 sec/word 68% correct</td>
<td>2.6s -0.2s 79% +11%</td>
<td>3.5s +0.7s 59% -9%</td>
<td>2.8s (+0.0s) 64% (-4%)</td>
<td>3.0s +0.2s 66% -2%</td>
</tr>
<tr>
<td>avg.</td>
<td>4.4 sec/word 47% correct</td>
<td>2.7s -1.7s 55% +8%</td>
<td>3.7s -0.7s 51% +4%</td>
<td>3.2s -1.2s 54% +7%</td>
<td>3.3s -1.1s 53% +6%</td>
</tr>
</tbody>
</table>
Unassisted Novice Translators

L1 = native French, L2 = native English, average time per input word only typing
Unassisted Novice Translators

L1 = native French, L2 = native English, average time per input word typing, initial and final pauses
Unassisted Novice Translators

L1 = native French, L2 = native English, average time per input word typing, initial and final pauses, short, medium, and long pauses

most time difference on intermediate pauses
## Activities: Native French User L1b

<table>
<thead>
<tr>
<th>User: L1b</th>
<th>total</th>
<th>init-p</th>
<th>end-p</th>
<th>short-p</th>
<th>mid-p</th>
<th>big-p</th>
<th>key</th>
<th>click</th>
<th>tab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unassisted</td>
<td>7.7s</td>
<td>1.3s</td>
<td>0.1s</td>
<td>0.3s</td>
<td>1.8s</td>
<td>1.9s</td>
<td>2.3s</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Postedit</td>
<td>4.5s</td>
<td>1.5s</td>
<td>0.4s</td>
<td>0.1s</td>
<td>1.0s</td>
<td>0.4s</td>
<td>1.1s</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Options</td>
<td>4.5s</td>
<td>0.6s</td>
<td>0.1s</td>
<td>0.4s</td>
<td>0.9s</td>
<td>0.7s</td>
<td>1.5s</td>
<td>0.4s</td>
<td>-</td>
</tr>
<tr>
<td>Prediction</td>
<td>2.7s</td>
<td>0.3s</td>
<td>0.3s</td>
<td>0.2s</td>
<td>0.7s</td>
<td>0.1s</td>
<td>0.6s</td>
<td>-</td>
<td>0.4s</td>
</tr>
<tr>
<td>Prediction+Options</td>
<td>4.8s</td>
<td>0.6s</td>
<td>0.4s</td>
<td>0.4s</td>
<td>1.3s</td>
<td>0.5s</td>
<td>0.9s</td>
<td>0.5s</td>
<td>0.2s</td>
</tr>
</tbody>
</table>
## Activities: Native French User L1b

<table>
<thead>
<tr>
<th>User: L1b</th>
<th>total</th>
<th>init-p</th>
<th>end-p</th>
<th>short-p</th>
<th>mid-p</th>
<th>big-p</th>
<th>key</th>
<th>click</th>
<th>tab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unassisted</td>
<td>7.7s</td>
<td>1.3s</td>
<td>0.1s</td>
<td>0.3s</td>
<td>1.8s</td>
<td>1.9s</td>
<td>2.3s</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Postedit</td>
<td>4.5s</td>
<td>1.5s</td>
<td>0.4s</td>
<td>0.1s</td>
<td>1.0s</td>
<td>0.4s</td>
<td>1.1s</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Options</td>
<td>4.5s</td>
<td>0.6s</td>
<td>0.1s</td>
<td>0.4s</td>
<td>0.9s</td>
<td>0.7s</td>
<td>1.5s</td>
<td>0.4s</td>
<td>-</td>
</tr>
<tr>
<td>Prediction</td>
<td>2.7s</td>
<td>0.3s</td>
<td>0.3s</td>
<td>0.2s</td>
<td>0.7s</td>
<td>0.1s</td>
<td>0.6s</td>
<td>-</td>
<td>0.4s</td>
</tr>
<tr>
<td>Prediction+Options</td>
<td>4.8s</td>
<td>0.6s</td>
<td>0.4s</td>
<td>0.4s</td>
<td>1.3s</td>
<td>0.5s</td>
<td>0.9s</td>
<td>0.5s</td>
<td>0.2s</td>
</tr>
</tbody>
</table>

Slightly less time spent on typing
# Activities: Native French User L1b

<table>
<thead>
<tr>
<th>User: L1b</th>
<th>total</th>
<th>init-p</th>
<th>end-p</th>
<th>short-p</th>
<th>mid-p</th>
<th>big-p</th>
<th>key</th>
<th>click</th>
<th>tab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unassisted</td>
<td>7.7s</td>
<td>1.3s</td>
<td>0.1s</td>
<td>0.3s</td>
<td>1.8s</td>
<td>1.9s</td>
<td>2.3s</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Postedit</td>
<td>4.5s</td>
<td>1.5s</td>
<td>0.4s</td>
<td>0.1s</td>
<td>1.0s</td>
<td>0.4s</td>
<td>1.1s</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Options</td>
<td>4.5s</td>
<td>0.6s</td>
<td>0.1s</td>
<td>0.4s</td>
<td>0.9s</td>
<td>0.7s</td>
<td>1.5s</td>
<td>0.4s</td>
<td>-</td>
</tr>
<tr>
<td>Prediction</td>
<td>2.7s</td>
<td>0.3s</td>
<td>0.3s</td>
<td>0.2s</td>
<td>0.7s</td>
<td>0.1s</td>
<td>0.6s</td>
<td>-</td>
<td>0.4s</td>
</tr>
<tr>
<td>Prediction+Options</td>
<td>4.8s</td>
<td>0.6s</td>
<td>0.4s</td>
<td>0.4s</td>
<td>1.3s</td>
<td>0.5s</td>
<td>0.9s</td>
<td>0.5s</td>
<td>0.2s</td>
</tr>
</tbody>
</table>

- Less pausing
- Slightly less time spent on typing
## Activities: Native French User L1b

<table>
<thead>
<tr>
<th>User: L1b</th>
<th>total</th>
<th>init-p</th>
<th>end-p</th>
<th>short-p</th>
<th>mid-p</th>
<th>big-p</th>
<th>key</th>
<th>click</th>
<th>tab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unassisted</td>
<td>7.7s</td>
<td>1.3s</td>
<td>0.1s</td>
<td>0.3s</td>
<td>1.8s</td>
<td>1.9s</td>
<td>2.3s</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Postedit</td>
<td>4.5s</td>
<td>1.5s</td>
<td>0.4s</td>
<td>0.1s</td>
<td>1.0s</td>
<td>0.4s</td>
<td>1.1s</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Options</td>
<td>4.5s</td>
<td>0.6s</td>
<td>0.1s</td>
<td>0.4s</td>
<td>0.9s</td>
<td>0.7s</td>
<td>1.5s</td>
<td>0.4s</td>
<td>-</td>
</tr>
<tr>
<td>Prediction</td>
<td>2.7s</td>
<td>0.3s</td>
<td>0.3s</td>
<td>0.2s</td>
<td>0.7s</td>
<td>0.1s</td>
<td>0.6s</td>
<td>-</td>
<td>0.4s</td>
</tr>
<tr>
<td>Prediction+Options</td>
<td>4.8s</td>
<td>0.6s</td>
<td>0.4s</td>
<td>0.4s</td>
<td>1.3s</td>
<td>0.5s</td>
<td>0.9s</td>
<td>0.5s</td>
<td>0.2s</td>
</tr>
</tbody>
</table>

- Less pausing
- Especially less time in big pauses
- Slightly less time spent on typing
### Origin of Characters: Native French L1b

<table>
<thead>
<tr>
<th>User: L1b</th>
<th>key</th>
<th>click</th>
<th>tab</th>
<th>mt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postedit</td>
<td>18%</td>
<td>-</td>
<td>-</td>
<td>81%</td>
</tr>
<tr>
<td>Options</td>
<td>59%</td>
<td>40%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Prediction</td>
<td>14%</td>
<td>-</td>
<td>85%</td>
<td>-</td>
</tr>
<tr>
<td>Prediction+Options</td>
<td>21%</td>
<td>44%</td>
<td>33%</td>
<td>-</td>
</tr>
</tbody>
</table>
Origin of Characters: Native French L1b

<table>
<thead>
<tr>
<th>User: L1b</th>
<th>key</th>
<th>click</th>
<th>tab</th>
<th>mt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postedit</td>
<td>18%</td>
<td>-</td>
<td>-</td>
<td>81%</td>
</tr>
<tr>
<td>Options</td>
<td>59%</td>
<td>40%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Prediction</td>
<td>14%</td>
<td>-</td>
<td>85%</td>
<td>-</td>
</tr>
<tr>
<td>Prediction+Options</td>
<td>21%</td>
<td>44%</td>
<td>33%</td>
<td>-</td>
</tr>
</tbody>
</table>

Translation comes to large degree from assistance
• Our classification of pauses is arbitrary (2-6sec, 6-60sec, >60sec)

• Extreme view: all you see is pauses
  – keystrokes take no observable time
  – all you see is pauses between action points

• Visualizing range of pauses:
  time $t$ spent in pauses $p \in P$ up to a certain length $l$

$$sum(t) = \frac{1}{Z} \sum_{p \in P, l(p) \leq t} l(p)$$
Results
Learning Effects

Users become better over time with assistance.
Learning Effects: Professional Translators

CASMACAT longitudinal study
Productivity projection as reflected in Kdur taking into account six weeks
(Kdur = user activity excluding pauses > 5 seconds)
Eye Trackers extensively used in cognitive studies of, e.g., reading behavior

- Overcomes weakness of key logger: what happens during pauses
- Fixation: where is the focus of the gaze
- Pupil dilation: indicates degree of concentration
Eye Tracking

- Problem: Accuracy and precision of gaze samples

\[ \times = \text{eye tracker result} \]
\[ \bullet = \text{target looked at} \]

Good precision, poor accuracy
Good accuracy, poor precision
Gaze-to-Word Mapping

- Recorded gaze locations and fixations

Left eye gaze samples

Families hit with increase in cost of living
British families have to cough up an extra £31,300 a year as food in supermarkets have climbed at an alarming rate over the past year, still, making it hard for the Bank of England to cut interest rates, control. To make matters worse, escalating prices are racing ahead healthcare professionals, who have suffered from the government below-inflation salary increases. In addition to fuel and food, elec

Right eye gaze samples

Gaze-to-word mapping

Families hit with increase in cost of living
British families have to cough up an extra £31,300 a year as food in supermarkets have climbed at an alarming rate over the past year, still, making it hard for the Bank of England to cut interest rates, control. To make matters worse, escalating prices are racing ahead healthcare professionals, who have suffered from the government below-inflation salary increases. In addition to fuel and food, elec
focus on target word (green) or source word (blue) at position $x$
• User style 1: Verifies translation just based on the target text, reads source text to fix it
User style 2: Reads source text first, then target text
Cognitive Studies: User Styles

- User style 3: Makes corrections based on target text only
• User style 4: As style 1, but also considers previous segment for corrections
## Users and User Styles

<table>
<thead>
<tr>
<th>Style 1</th>
<th>Style 2</th>
<th>Style 3</th>
<th>Style 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>target / source-fix</strong></td>
<td><strong>source-target</strong></td>
<td><strong>target only</strong></td>
<td><strong>wider context</strong></td>
</tr>
<tr>
<td>P</td>
<td>PI</td>
<td>PIA</td>
<td>P</td>
</tr>
<tr>
<td>P02</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>P03</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>P04</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>P05</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>P07</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>P08</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>P09</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
</tbody>
</table>

- Individual users employ different user styles
- But: consistently across different types of assistance
  (P = post-editing, PI = interactive post-editing, PIA = interactive post-editing with additional annotations)
Backtracking

• Local backtracking
  – **Immediate repetition**: the user immediately returns to the same segment (e.g. AAAA)
  – **Local alternation**: user switches between adjacent segments, often singly (e.g. ABAB) but also for longer stretches (e.g. ABC-ABC).
  – **Local orientation**: very brief reading of a number of segments, then returning to each one and editing them (e.g. ABCDE-ABCDE).

• Long-distance backtracking
  – **Long-distance alternation**: user switches between the current segment and different previous segments (e.g. JCJDJFJG)
  – **Text final backtracking**: user backtracks to specific segments after having edited all the segments at least once
  – **In-text long distance backtracking**: instances of long distance backtracking as the user proceeds in order through the text.
Thank You

questions?