Advances in Computer Aided Translation Beyond Post-Editing

Philipp Koehn

20 May 2016
• Screenshot from CASMACAT post-editing mode (same as MATECAT)

• Source on left, translation on right / context above and below
Productivity Improvements

(source: Autodesk)
Le Pakistan a donc été récompensé par l’assistance et les armes des États-Unis. As a result, Pakistan was rewarded with American financial assistance and arms.

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.

- Caret alignment (green)
- Mouse alignment (yellow)
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
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
Interactive Translation Prediction

Forget it. It's too risky. I'm through doing that shit.

Olvidarlo. Es demasiado arriesgado. Estoy haciendo

You always say that. The same thing every time.

"I'm through, never again, too dangerous."
Shade Off Translated

- Word alignment visualization for interactive translation prediction
- Shade off words that are already translated
- Highlight words aligned to first predicted translation word
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?
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
• European research project 2011-2014

• All describe methods implemented in CASMACAT workbench
  – builds on MATECAT open source implementation
  – typical web application: LAMP (Linux, Apache, MySQL, PHP)
  – uses model, view, controller breakdown

• Workbench freely available at http://www.casmacat.eu/
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
Projects at the MT Marathon

- Implementation of word level confidence measures (Winston Wu)
- Improvements to bilingual concordancer (Winston Wu)
- De-duplication of translation option array (Rebecca Marvin)
- More configuration options (Justin Debenedetto)
- Debugging of replay mode (Justin Debenedetto)
- Pre-post-editing (Lane Schwartz and Philipp Koehn)
Thank You

questions?