SECTra_w.1 : an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora

Cong-Phap HUYNH
Christian BOITET
Hervé BLANCHON

Laboratoire LIG, GETALP, GETA
385, rue de la Bibliothèque, 38041 Grenoble, France
Outline

- Introduction
- Motivations
  - make multilingual parallel corpora
    - richer: multimedia (text, speech), multiannotated, multiusage
    - accessible through the Web (replay, reading)
    - processable through the Web (evaluation, annotation, improvement, extension to other languages)
- SECTra_w.1 is mainly for:
  - storing & presenting parallel multimedia corpora
  - evaluating of MT systems
    - classical evaluations (MT evaluation campaigns)
      - subjective (fluidity, adequacy)
      - objective (n-gram-based metrics such as BLEU, NIST...)
    - task-related objective evaluation
Introduction

Recent MT evaluation campaigns have been criticized
- only tables of figures such as scores (BLEU, NIST, ORANGE, METEOR…) are shown as results
- these n-gram-based measures have been shown not to correlate very well with human judgments, contrary to initial expectations

[Callison-Burch & al. 2006]

Commercial MT systems are ranked
- consistently low by these measures
- quite high, or highest by human judges

Hence, it would be good to look at the real data and to "see for ourselves"

SECTra_w.1
- shows tables of figures as results of an evaluation campaign
- shows the real data (source, MT outputs, references, post-edited outputs)
- makes the post-editing effort sensible
Motivations

- Subjective evaluation biases
  - showing a reference translation of an input instead of the input itself when judging adequacy
    - Reference translations are not 100% adequate!
  - showing to 1 judge the results of several MT systems in parallel
    - S/he compares them instead of grading them independently

- Other people would like to do subjective evaluation after the evaluation campaign
  - in a certain application setting
  - for their research

- Proposal: task-related measure of the "real" MT quality
  - i.e. utility of MT outputs for a certain task
  - Tasks: HQ understanding, HQ translation, HQ communication
  - SECTra_w.1 concentrates now on producing HQ translations
    - MT post-editing
    - measures of the effort
SECTRa_w.1 used for MT evaluation

Dc: character distance
Dw: word distance
D: sentence distance


Fluency: F1: written, F2: oral, F3: not acceptable

<table>
<thead>
<tr>
<th>Source</th>
<th>Translation</th>
<th>Distance</th>
<th>BLEU</th>
<th>NIST</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>About ten minutes.</td>
<td>Approximately dix minutes.</td>
<td>Dc=0, Dw=0</td>
<td>1.0</td>
<td>2.0</td>
<td>Approximativement dix minutes.</td>
</tr>
<tr>
<td>Actually I'm on my period.</td>
<td>Résolament je suis sur ma période.</td>
<td>Dc=28, Dw=6</td>
<td>0.61</td>
<td>0.37</td>
<td>En fait, réellement je suis dans ma période.</td>
</tr>
<tr>
<td>Is that a problem?</td>
<td>Est-ce que c'est un problème?</td>
<td>Dc=1, Dw=2</td>
<td>0.65</td>
<td>1.8</td>
<td>Est-ce que c'est un problème?</td>
</tr>
<tr>
<td>All right.</td>
<td>Tout le droit.</td>
<td>Dc=12, Dw=3</td>
<td>0.71</td>
<td>0.25</td>
<td>D'accord, tout le droit.</td>
</tr>
<tr>
<td>It's two thirty now, so you're supposed to</td>
<td>C'est maintenant deux trente, donc vous êtes</td>
<td>Dc=34, Dw=6</td>
<td>0.35</td>
<td>3.69</td>
<td>Ambulance, allez vous plait.</td>
</tr>
<tr>
<td>Ambulance, please.</td>
<td>Ambulance, allez-vous plait.</td>
<td>Dc=0, Dw=0</td>
<td>1.0</td>
<td>2.68</td>
<td>Ambulance, allez vous plait.</td>
</tr>
</tbody>
</table>
Classical MT evaluation in SECTra_w

- Subjective evaluation
  - Principles and organization
    - Several judges can perform evaluation
      - at the same time on the same part of the data
    - A segment receives several evaluation scores
      - these scores can be seen by users
    - Integration of a "workflow"
      - to define the judges and assign them sets of pages to evaluate
  - Protocol
    - slightly different of the protocol proposed by NIST
      - In TIDES evaluation campaign
    - Fluency
      - F1: written, F2: oral, F3: not acceptable
    - Adequation
Classical MT evaluation in SECTra\textsubscript{w} (2)

- **Objective n-gram-based scores**
  - n-gram-based measures
    - we started from NIST scripts for BLEU, NIST…
    - reimplementation in Java, integration in SECTra\textsubscript{w}
  - **Edit distances**
    - Character, word, and sentence distances
      - **Character:** $D_c = $ Levenshtein distance (Wagner&Fisher)
        - Cost $(c1, c2) = 1$
      - **Word:** $D_w = $ pseudo-distance (Damereau & Fisher)
        - Cost $(w1, w2) = D_c(w1, w2)$
        - $D_w(s1, s2) = D'_c(s1, s2)$ with 'characters' = words
      - **WER=$D_w$ and character-based distance=$D_c$**
    - **Sentence distance**
      - $D_{Sent} = \alpha \ D_c + \beta \ D_w$ \hfill ($\alpha$ and $\beta$ modifiable, sum to 1)
Task-related Objective Evaluation

- Time taken to perform the task: post-edition, booking, understanding…
  - In the profession, translators are paid by word or by page
    - with rates corresponding to the time taken
    - itself linked to the difficulty of the task
  - The simplest and most reliable measure is the sentence distance
    - informative if one measures $D(\text{MT result, MT post-edited})$
    - usable to deduce the post-edition time even if done offline
  - Measures such as WER and mWER, as currently used,
    - are edit distances,
    - but are not related to the task performed
    - because they compare an a priori reference and the result
<table>
<thead>
<tr>
<th>Source</th>
<th>Translation (Reverso)</th>
<th>Distance</th>
<th>Post-edition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamburger and stew on the right side and salad, please.</td>
<td>Hamburger et ragoût à droite côté et salade, s'il vous plaît.</td>
<td>Dc=20, Dw=7</td>
<td>Un Hamburger et du ragoût à droite sur le côté et de la salade, s'il vous plaît.</td>
</tr>
<tr>
<td></td>
<td>Dsent=9.6</td>
<td></td>
<td>(A2)</td>
</tr>
<tr>
<td>That fried fish, one sausage with green peas, please.</td>
<td>Cela a frit du poisson, une saucisse avec les pois verts, s'il vous plaît.</td>
<td>Dc=25, Dw=3</td>
<td>Ce poisson frit, Cela a frit du poisson, une saucisse avec les pois petits verts, pois, s'il vous plaît.</td>
</tr>
<tr>
<td></td>
<td>Dsent=11.4</td>
<td></td>
<td>(A3)</td>
</tr>
<tr>
<td>T-bone steak and sauerkraut and fried potatoes, please.</td>
<td>Steak avec un os en T et choucroute et a frit des pommes de terre, s'il vous plaît.</td>
<td>Dc=33, Dw=11</td>
<td>Du bifteck à l'os steak et avec de un la os en T et choucroute et a frit des pommes de terre, terre frites, s'il vous plaît.</td>
</tr>
<tr>
<td></td>
<td>Dsent=15.4</td>
<td></td>
<td>(A3)</td>
</tr>
<tr>
<td>Roast chicken and two slices of ham on this side and spinach, please.</td>
<td>Poulet du rôti et deux tranches de jambon sur ce côté et épinards, s'il vous plaît.</td>
<td>Dc=8, Dw=2</td>
<td>Du Poulet du rôti et deux tranches de jambon sur ce côté et des épinards, s'il vous plaît.</td>
</tr>
<tr>
<td></td>
<td>Dsent=3.2</td>
<td></td>
<td>(A2)</td>
</tr>
<tr>
<td>I'd like breakfast, please.</td>
<td>J'aimerais petit déjeuner, s'il vous plaît.</td>
<td>Dc=3, Dw=1</td>
<td>J'aimerais un J'aimerais petit déjeuner, s'il vous plaît.</td>
</tr>
<tr>
<td></td>
<td>Dsent=1.4</td>
<td></td>
<td>(A1)</td>
</tr>
</tbody>
</table>
Interface of task-related objective evaluation

<table>
<thead>
<tr>
<th>Source (English)</th>
<th>Post-Edition (French) (170/502=33.86%)</th>
<th>Possibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Its lower border is the level of the ocean</td>
<td>Son cadre inférieur est le niveau de l’océan</td>
<td><strong>Syslan</strong></td>
</tr>
<tr>
<td>The turnover time is tens and hundreds of thousands of years</td>
<td>Le temps de rotation est des dizaines et des centaines de milliers d’années</td>
<td><strong>Syslan</strong></td>
</tr>
<tr>
<td>The lower zone is a zone of difficult water exchange</td>
<td>La zone inférieure est une zone d’échange difficile de l’eau</td>
<td><strong>Syslan</strong></td>
</tr>
<tr>
<td>It is located below the ocean level</td>
<td>Elle est située au-dessous du niveau d’océan</td>
<td><strong>Syslan</strong></td>
</tr>
</tbody>
</table>
Experiment and validation

- **Evaluation campaign**
  - funded by FT R&D (TRANSAT project)
  - 2 English-French corpora
    - constructed from English segments from the BTEC corpus
    - In total ~5000 evaluation sentences
    - distributed by ATR for IWSLT-06
      - 2 commercial MT systems and 1 research MT system
      - post-editions of MT results
  - subjective evaluation
    - 12 judges
    - subjective evaluation cost compared with that of IWSLT-06
      - time diminished by a factor of about 5
Conclusion

- **SECTra_w.1**
  - a web-oriented system for managing multilingual corpora
  - mainly dedicated to the evaluation of MT systems
    - subjective evaluation (fluidity, adequacy)
    - objective n-gram-based scores
    - task-oriented evaluation
      ♦ by letting humans post-edit the MT results
      ♦ measuring an edit distance (and/or the post-editing time)
  - **Experiment**
    - In an evaluation campaign funded by FT R&D (Nov. 2008)
    - on two English-French corpora
      ♦ 5,000 sentences translated by Systran and Reverso
      ♦ 12 judges
  - **Conclusions**
    - Good usability, practical, reliable, fast enough
    - Still many improvements in view
Future work

- Improve user and work management
  - Put workflow under control of linguistic managers
  - Enable direct communication between users
    - Now only possible by posting comments (or e-mail...)
    - Inspiration from BEYRans (Y. Bey) — for volunteer translators

- SECTra_w.2 (already in operation)
  - Used to support post-editing in the EOLSS/UNL-fr project
    - 25 documents, \( \approx 220 \text{K words or 880 pages} \)
    - \( \approx 25 \) volunteer (researchers) and paid (students) translators
    - Reverso, Systran, UNL-fr deconversion for pretranslation
  - Will be included as component of various iMAGs
    - iMAG = interactive Multilingual Access Gateway
      - Translation gateway dedicated to some web site/domain
        - Having a specialized TM (SECTra_w) and LexDB
Related publications


The end

Cảm ơn
Thank you
Merci