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-edition 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

| Source                          | Translation                          | Distance | BLEU | NIST | Reference                        |
|--------------------------------|--------------------------------------|----------|------|------|----------------------------------|
| About ten minutes.             | Approximately dix minutes.           | Dc=0, Dw=0, D=0.0 | 1.0  |      | Approximately dix minutes.       |
| Actually I'm on my period.     | Réellement je suis sur ma période.   | Dc=26, Dw=6, D=12.9 | 0.61 | 0.37 | En fait, Réséllent j ai mox suer portéla, sur ma péroda. |
| Is that a problem?             | Est-ce que c'est un problème?        | Dc=1, Dw=2, D=1.7  | 0.55 | 1.8  | Est-ce que c'est un problème ?    |
| All right.                     | Tout le droit.                       | Dc=12, Dw=3, D=5.7 | 0.71 | 0.25 | D'accord, Tout le droit.          |
| It's two thirty now, so you're | C'est maintenant deux trente, donc vous êtes | Dc=34, Dw=6, D=18.5 | 0.35 | 3.69 |                                |
| supposed to                   |                                     |           |      |      |                                  |

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

**Adequation:** A1 : All, A2 : Almost all, A3 : Half, A4 : Few, A5 : None

**Distance:**
- Dc: character distance
- Dw: word distance
- D: sentence distance

©Huynh C.P., Ch. Boitet, H. Blanchon
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
      - A1 : All, A2 : Almost all, A3: Half, A4 : Few, A5 : None
Classical MT evaluation in SECTra_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_w
  - Edit distances
    - Character, word, and sentence distances
      - Character: $D_c = \text{Levenshtein distance (Wagner}&Fisher)$
        - Cost $(c1, c2) = 1$
      - Word: $D_w = \text{pseudo-distance (Damereau}& &Fisher)$
        - Cost $(w1, w2) = D_c (w1, w2)$
        - $D_w (s1, s2) = D'_c (s1, s2) \text{ with 'characters' = words}$
      - WER=$D_w$ and character-based distance=$D_c$
    - Sentence distance
      - $D_{Sent} = \alpha D_c + \beta D_w$ ($\alpha$ 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
## Evaluation results and post-edition effort

| Source | Translation (Reverso) | Distance | Post-edition | HERVE | GEORGES | ACHILLE |
|--------|-----------------------|----------|--------------|-------|----------|---------|
| Hamburger and stew on the right side and salad, please. | Hamburger et ragoût à droite côté et salade, s'il vous plaît. | Dc=20, Dw=7, Dsent= 9.6 | Un Hamburger et du ragoût à droite sur le côté et de la salade, s'il vous plait. | (A2) | (A2) | (A1) |
| That fried fish, one sausage with green peas, please. | Cela a frit du poisson, une saucisse avec les pois verts, s'il vous plaît. | Dc=25, Dw=3, Dsent= 11.4 | Ce poisson frit, Cela a frit du poisson, une saucisse avec les des pois petits verts, pois, s'il vous plaît. | (A3) | (A2) | (A2) |
| T-bone steak and sauerkraut and fried potatoes, please. | Steak avec un os en T et choucroute et a frit des pommes de terre, s'il vous plaît. | Dc=33, Dw=11, Dsent= 15.4 | Du bifteck à l'os steak et avec de un la os en T et choucroute et « frit des pommes de terre, terre frites, s'il vous plaît. | (A3) | (A2) | (A3) |
| Roast chicken and two slices of ham on this side and spinach, please. | Poulet du rôti et deux tranches de jambon sur ce côté et épinards, s'il vous plaît. | Dc=8, Dw=2, Dsent= 3.2 | Du Poulet du rôti et deux tranches de jambon sur ce côté et des épinards, s'il vous plaît. | (A2) | (A2) | (A2) |
| I'd like breakfast, please. | J'aimerais petit déjeuner, s'il vous plait. | Dc=3, Dw=1, Dsent= 1.4 | J'aimerais un J'aimerais petit déjeuner, s'il vous plait. | (A1) | (A1) | (A1) |
## Interface of task-related objective evaluation

| ID | Source (English) | Post-Editing (French) | Possibilities |
|----|------------------|----------------------|---------------|
| 201 | The turnover time is tens and hundreds of thousands of years | Le temps de rotation est des dizaines et des centaines de milliers d’années | **Syrax**<br>**Reverse**<br>The temps de rotation est des dizaines et des centaines de milliers d’années<br>The temps de rotation est des dizaines et des centaines de milliers d’années |
| 200 | It is located below the ocean level | Elle est située au-dessous du niveau d’océan | **Syrax**<br>**Reverse**<br>It est recherché au-dessous du niveau d’océan<br>It est recherché au-dessous du niveau d’océan |
| 200 | The lower zone is a zone of difficult water exchange | La zone inférieure est une zone d’échange difficile de l’eau | **Syrax**<br>**Reverse**<br>The la zone inférieure est une zone d’échange de l’eau difficile<br>The la zone inférieure est une zone d’échange de l’eau difficile |
| 200 | Its lower border is the level of the ocean | Son cadre inférieur est le niveau de l’océan | **Syrax**<br>**Reverse**<br>Its la bordure inférieure est le niveau de l’océan<br>Its la bordure inférieure est le niveau de l’océan |

---

**The:** la [preposition-other] le [preposition-other] les [personal pronoun]  
**time:** de [nominative] temps [nominative]  
**is:** est [verb] fait [verb]  
**and:** et [conditional]  
**hundreds:** centaines [noun-plural]  
**of:** de [preposition-other]
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-edition in the EOLSS/UNL-fr project
    - 25 documents, ≈ 220K words or 880 pages
    - ≈ 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

- Huynh C. P., Boitet Ch., Fafiotte G. (2008) *Extending an On-line Parallel Corpus Management System to Handle Specific Types of Structured Documents*. SLTU Conference, MICA, Hanoi, Viet Nam, May 2008, 6p.

- Huynh C. P., Boitet Ch., Blanchon H., Nguyen H. T. (2008) *A Web-oriented System to Manage the Translation of an Online Encyclopedia Using Classical MT and Deconversion from UNL*. COLING Conference, Manchester, England, August 2008, 7p. (submitted).
The end

Cảm ơn
Thank you
Merci