Experience and Conclusions from the CESTA Evaluation Project

Olivier Hamon

ELDA, France
Outline

1. Introduction

2. Automatic metrics used

3. CESTA experience

4. Open issues & conclusions
Introduction

CESTA:
- Two Evaluation campaigns of machine translation systems
- 13 different systems
- Arabic-to-French and English-to-French directions
- Observe the behaviour of well-known metrics for those directions
- Experiment with new metrics
- Conduct a meta-evaluation

Automatic metrics used within CESTA
Automatic metrics used

Widely-used and well-known metrics:

- **BLEU**: *Bilingual Evaluation Understudy* (Papineni et al., 2001)
  - Weighted average of common n-grams between the hypothesis and the references
  - Needs 1..n references (CESTA=4)
  - Good reliability in previous experiments

- **NIST**: (Doddington, 2002)
  - Like BLEU but considers information gain and length penalty
  - Needs 1..n references (CESTA=4)
  - Outperforms BLEU in previous experiments
Automatic metrics used

- WNM: *Weighted N-gram Metric* (Babych & Hartley, 2004)
  - Combines BLEU with weight of statistical salience
  - Needs 1 reference (CESTA=1 to 4) and a statistical corpus
  - Outperforms BLEU and NIST in previous experiments
Automatic metrics used

Experimental metrics:

– **X-Score (Rajman & Hartley, 2001)**
  - Analysis of the grammaticality of the hypothesis. The morpho-syntactical distribution is compared with a reference corpus fluency-annotated
  - Needs a fluency-annotated corpus

– **D-Score (Rajman & Hartley, 2001)**
  - Analysis of the preservation of the semantic content between the source and the hypothesis. The semantic vector model of the hypothesis is compared with a reference
  - Needs a parallel corpus
CESTA experience
CESTA experience – BLEU / NIST

– NIST correlation slightly better than BLEU correlation

– But it is « easier » to understand BLEU (scale 0-100) than NIST (no scale)

– BLEU and NIST correlations not as good as expected
Open issues & conclusions

Amount of reference translations (BLEU)

![Graph showing BLEU scores for different runs with legend: RUN1_moy, RUN2_moy, RUN2-OOD_moy]
CESTA experience – WNM

- Adaptation to the NIST format for CESTA
- Much better correlation than BLEU / NIST
- Correlation dependant on the references

<table>
<thead>
<tr>
<th>2nd run En→Fr</th>
<th>Ref-1</th>
<th>Ref-2</th>
<th>Ref-3</th>
<th>Ref-4</th>
<th>Comb.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>83.19</td>
<td>86.16</td>
<td>96.73</td>
<td>83.94</td>
<td>85.58</td>
</tr>
<tr>
<td>Adequacy</td>
<td>94.23</td>
<td>94.86</td>
<td>87.78</td>
<td>94.16</td>
<td>95.11</td>
</tr>
</tbody>
</table>

→Needs reference combination as BLEU does? (‘mWNM’)
CESTA experience – X-Score

- Only the translation is considered for the metric
- The translation is characterized by the occurrence frequency profile of syntactic features (POS tags in our case)
- The frequency profiles are used to train a linear predictor for the fluency score
- Two stages:
  - Learning phase: production of the grammaticality model (i.e. computation of the coefficients of the linear predictor)
  - Evaluation phase: computation of the scores
CESTA experience – X-Score

1. Training corpus (=fluency corpus)

2. POS tagging

3. POS tag occurrence frequency profile

4. Linear predictor

5. Association of frequency profile / fluency scores

Hypothesis

1. POS tagging

2. POS tag occurrence frequency profile

3. Application of the linear predictor

4. X-Score = predicted fluency score

Fluency score (human judges)
CESTA experience – X-Score

- Not correlated

⇒ Reconsidering the problem

- Several issues are raised:
  - Tagger dependent
  - Weights are too high and favour some tags ⇒ a solution is to compute the ratio of tags
  - Word ordering ⇒ needs to use n-grams, but very time consuming (CESTA : 35 tags, 1,156 bi-grams, 1M ratio, resulting a 1B entry matrix!)
  - Selection of tags
  - ...

CESTA experience – D-Score

• Hypothesis: source and target languages have the same semantic vector. Similarity comparison between documents

• Use of a large parallel corpus

• Two stages:
  – Learning phase:
    • For the whole corpus, computing of the relative term-frequency vectors in document
    • For each document, computing the relative document-frequency vectors in terms
    • Each parallel document has a position in its language vector space
  – Evaluation phase: Computing of similarities with each document of the corpus, for source and target documents
CESTA experience – D-Score

Source document

\[ \text{Similarity vector } V \]

\[ D_1 \quad D_2 \quad \ldots \quad D_N \]

Target document

\[ \text{Similarity vector } V' \]

\[ D'_1 \quad D'_2 \quad \ldots \quad D'_N \]

Source corpus

Target corpus

Computing distance

D-Score
CESTA experience – D-Score

– Correlations are inconsistent
– Need to be studied in depth (ongoing)

➤ Maybe reconsidering the problem?

– A lot of parameters (filtering, which tags, tagger, etc.)
Open issues & conclusions
Open issues & conclusions

- Reliability of BLEU / NIST, WNM corresponds to literature
- For BLEU, NIST, WNM, fluency correlations slightly higher than adequacy correlations; except on a specific domain (vocabulary)
- Bad correlations for X-Score, D-Score
- Experimental metrics not ready yet
- Task / domain dependant
Open issues & conclusions

– Do we need so many metrics?

– BLEU, NIST, WNM, etc.:
  • Obtain similar same correlations most of the time
  • Give the same analysis: are the hypothesis words present in the references? In correct order?

– other metrics, but computing other things? (that do not rely with n-grams...)
Open issues & conclusions

- Costs (money and time) for CESTA:
  - BLEU / NIST / WNM = reference translations
    ~ 4 * 2,000€ (cannot be reduced)
    ~ 2/3 weeks (not easy to reduce)
  - X-Score = reference corpus
    ~ 38 * 30€ (could be reduced)
    ~ 3/4 weeks (could be reduced)
  - D-Score = parallel corpus
    ~ 0 (already available), but very large cost
  - Human = judges
    ~ 100 * 30€ (for the first campaign)
    ~ 3/4 weeks
Open issues & conclusions

– Is it really cheaper to use automatic metrics instead of human evaluation?
  • for a single campaign → not really
  • for systems → yes?
  • data evolve quickly…
  • less data also allows to know systems’ quality