QUALITY ESTIMATION AND EVALUATION OF MACHINE TRANSLATION INTO ARABIC

Houda Bouamor, Carnegie Mellon University-Qatar
email: hbouamor@qatar.cmu.edu
The WikiTEA project at CMU-Q

The Arabic content in Wikipedia is less than needed

Welcome to Wikipedia, the free encyclopedia that anyone can edit.
4,343,512 articles in English

Arabic rank: 27 out of 280 languages

Automatically create an Arabic summary for a given English article
Outline

1. SUMT: A Framework of Summarization and MT
2. AL-BLEU: Metric and a dataset for Arabic MT evaluation
3. Conclusions
SUMT: A Framework of Summarization and Machine Translation
Motivation

- MT quality is far from ideal for many languages and text genres
  - Provides incorrect context and confuses readers
- Some of sentences are not as informative
  - Could be summarized to make a more cohesive document

Keep informative sentences + decent MT quality
Questions?

1. How can we estimate the MT quality of a sentence without human references?
2. How can we find the most informative part of a document?
3. How can we find a middle point between informativeness and MT quality?
4. How can we evaluate the quality of our system?
Part 1: outline

1. MT quality estimation
2. MT-aware summarization system
3. Experiments and results
4. Conclusion
SuMT [Bouamor et al., 2013]
SuMT: Translation

Build a standard English to Arabic Machine translation system
SuMT: MT quality estimation

Design a binary classifier: \( <\text{high} ; \text{low} > \)
SuMT: MT quality estimation

Data labeling procedure

(a) Measure the TER score [Snover et al., 2006] of Doc
(b) Measure TER(Sent)
   - If \( \text{TER(Sent)} > \text{TER(Doc)} \): Sent has low translation quality.
   - Else: Sent has high translation quality.

\[\begin{align*}
\text{Sent}^{\text{EN}}_1, \text{Sent}^{\text{AR}}_1 & : \text{Q Score}_1 \\
\text{Sent}^{\text{EN}}_2, \text{Sent}^{\text{AR}}_2 & : \text{Q Score}_2 \\
\text{Sent}^{\text{EN}}_3, \text{Sent}^{\text{AR}}_3 & : \text{Q Score}_3 \\
\vdots & \\
\text{Sent}^{\text{EN}}_n, \text{Sent}^{\text{AR}}_n & : \text{Q Score}_n
\end{align*}\]

\(\text{Sent}^{\text{EN}}_n\): A source sentence
\(\text{Sent}^{\text{AR}}_n\): Its automatically obtained translation
Quality Estimation: MT Quality classifier

- Use SVM classifier
- Adapt Quest framework [Specia et al., 2013] to our EN-AR translation setup
- Each sentence is characterized with:
  - General features: length, ratio of S-T length, S-T punctuations
  - 5-gram LM scores
  - MT-based scores
  - Morphosyntactic features
  - …
SuMT: MT-aware summarization

Select most informative + higher MT quality sentences
SuMT: MT-aware summarization

MEAD as a ranker

\[ \text{Rank}(\text{Sent}_i^{EN}) = \alpha \times \text{position}(\text{Sent}_i^{EN}) + \beta \times \text{centroid}(\text{Sent}_i^{EN}) + \lambda \times \text{length}(\text{Sent}_i^{EN}) \]
SuMT: MT-aware summarization

Our adaptation of MEAD

\[
\text{Rank}'(\text{Sent}_i^{EN}) = \alpha \times \text{position}(\text{Sent}_i^{EN}) + \beta \times \text{centroid}(\text{Sent}_i^{EN}) + \lambda \times \text{length}(\text{Sent}_i^{EN}) + \gamma \times Q_{\text{score}i}
\]
Evaluation quality estimation

- How do we evaluate the quality of the estimation?
  - Intrinsically: very hard to trust
    - Need references $\rightarrow$ MT evaluation
    - Next...
  - Extrinsically: in an application
    - In the context of MT of Wikipedia
    - Compare using QE vs. a simple baseline
SuMT: experimental settings

- **MT setup**
  - Baseline MT system: MOSES trained on a standard English-Arabic corpus
  - Standard preprocessing and tokenization for both English and Arabic
  - Word-alignment using GIZA++

- **Summarization and test data**
  - English-Arabic NIST corpora
    - Train: NIST 2008 and 2009 for the training and development (259 documents)
    - Test: NIST 2005 (100 documents)
SuMT: experimental settings

- **Summarization setup**
  - Bilingual summarization of the test data
  - 2 native speakers chose half of the sentences
  - **Guidelines in sentence selection:**
    - Being informative in respect to the main story
    - Preserving key informations (NE, dates, etc.)
  - A moderate agreement of $K=0.61$.
SuMT: experimental settings

- Producing summaries for each document using:
  - Length-based: choose the shortest sentences (Length)
  - State of the art MEAD summarizer (MEAD)
  - MT quality estimation classifier (Classifier)
  - MT-aware summarizer (SuMT)
  - Oracle classifier: choose the sentences with the highest translation quality (Oracle).
SuMT: Results

MT results

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>27.52</td>
</tr>
<tr>
<td>Length</td>
<td>26.33</td>
</tr>
<tr>
<td>MEAD</td>
<td>28.42</td>
</tr>
<tr>
<td>Classifier</td>
<td>31.36</td>
</tr>
<tr>
<td>Interpol</td>
<td>28.45</td>
</tr>
<tr>
<td>SuMT</td>
<td>32.12</td>
</tr>
<tr>
<td>Oracle</td>
<td>34.75</td>
</tr>
</tbody>
</table>

Bouamor et al. 2013: SuMT 19 / 26
SuMT: Results

Arabic Summary quality

<table>
<thead>
<tr>
<th>Method</th>
<th>Rouge-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>15.81</td>
</tr>
<tr>
<td>MEAD</td>
<td>23.56</td>
</tr>
<tr>
<td>Classifier</td>
<td>23.09</td>
</tr>
<tr>
<td>Interpol</td>
<td>20.33</td>
</tr>
<tr>
<td>SuMT</td>
<td>24.07</td>
</tr>
</tbody>
</table>
Conclusions

- Presented a framework for pairing MT with summarization
- We extend a classification framework for reference-free prediction of translation quality at the sentence-level.
- We incorporate MT knowledge into a summarization system which results in high quality translation summaries.
- Quality estimation is shown to be useful in the context of text summarization.
Automatic MT quality evaluation
The BLEU metric

- De facto metric, proposed by IBM [Papineni et al., 2002]

- Main ideas
  - Exact matches of words
  - Match against a set of reference translations for greater variety of expressions
  - Account for adequacy by looking at word precision
  - Account for fluency by calculating n-gram precisions for n=1,2,3,4
  - No recall: difficult with multiple refs: “Brevity penalty”, instead
  - Final score: a weighted geometric average of the n-gram scores
The BLEU metric: Example

• Example:
  – Reference: “the Iraqi weapons are to be handed over to the army within two weeks”
  – MT output: “in two weeks Iraq’s weapons will give army”

• BLUE metric:
  – 1-gram precision: 4/8
  – 2-gram precision: 1/7
  – 3-gram precision: 0/6
  – 4-gram precision: 0/5
The BLEU metric: Example

Example:

- Reference: “the Iraqi weapons are to be handed over to the army within two weeks”
- MT output: “in two weeks Iraq’s weapons will give army”

BLUE metric:

- 1-gram precision: 4/8
- 2-gram precision: 1/7
- 3-gram precision: 0/6
- 4-gram precision: 0/5

This example is taken from Alon Lavie’s AMTA 2010 MT evaluation tutorial
The BLEU metric: Example

• Example:
  – Reference: “the Iraqi weapons are to be handed over to the army within two weeks”
  – MT output: “in two weeks Iraq’s weapons will give army”

• BLEU metric:
  – 1-gram precision: 4/8
  – 2-gram precision: 1/7
  – 3-gram precision: 0/6
  – 4-gram precision: 0/5

\[
\text{BLEU} \approx \text{smoothing} \left( (\prod P_{i\text{-gram}})^{1/n} \times \text{ (brevity penalty) } \right)
\]

Short story: the score is [0..1]; higher is better!
BLEU & Arabic

- **BLEU heavily penalizes Arabic**

<table>
<thead>
<tr>
<th>Source</th>
<th>Reference</th>
<th>Rank&lt;sub&gt;BLEU&lt;/sub&gt;</th>
<th>Rank&lt;sub&gt;Human&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>France plans to attend ASEAN emergency summit.</td>
<td>فرنسا تعتزم حضور قمة الآسيان الطارئة.</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MT translation (1)</td>
<td>فرنسا تخطط لحضور القمة الطارئة للأسيا</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MT translation (2)</td>
<td>فرنسا لحضور قمة الآسيان خطط الطوارئ</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

- **Question**: How can we adapt BLEU to support Arabic Morphology?
For our experiments:

1. AL-BLEU metric
2. Data and systems
AL-BLEU: Arabic Language BLEU

- Extend BLEU to deal with Arabic rich morphology
- Update the n-gram scores with **partial credits** for partial matches:
  - **Morphological**: POS, gender, number, person, definiteness
  - **Stem and lexical matches**
- Compute a new matching score as follows:

\[
match(t_h, t_r) = \begin{cases} 
1, & \text{if } t_h = t_r \\
ws + \sum_{i=1}^{5} wf_i & \text{otherwise}
\end{cases}
\]

- \(t_h\): hypothesis token
- \(t_r\): reference token
- \(ws\): stem weight
- \(wf_i\): morph. weights
AL-BLEU: Arabic Language BLEU

**HYP:** فرنسا تخطط لحضور القمة الطارئة للأسيان

**REF:** فرنسا تعترف بحضور قمة الأسيان الطارئة

- exact
- stem & morph.
- stem only
- morph. only
AL-BLEU: Arabic Language BLEU

- MADA [Habash et al., 2009] provides stem and morph. Features
- Weights are optimized towards improvement of correlation with human judgments
- Hill climbing used on development set
- AL-BLEU is a geometric mean of the different matched n-grams
A good MT metric should correlate well with human judgments.

Measure the correlation between BLEU, AL-BLEU and human judgments at the sentence level.
Problem: “No” human judgment dataset for Arabic

Data

- Annotate a corpus composed of different text genres:

Systems

- Six state-of-the-art EN-AR MT systems
  - 4 research-oriented systems
  - 2 commercial off-the-shelf systems
Data: Judgment collection

- Rank the sentences relatively to each other from the best to the worst

Political relations between the two countries have been cool in the past because of an exchange of accusations on the matter of supporting terrorism. Moscow asserted that Turkey turned a blind eye to Chechen activists on its soil while Ankara accuses Russia of harboring Kurdish separatists. Putin stressed that "there are common points in the standpoint on how to fight terrorism." The joint statement issued by the two presidents particularly mentions the importance of "reinforcing joint efforts in the area of fighting terrorism" as well as "condemning terrorism in all its forms."

Source
Rank the sentences relatively to each other from the best to the worst

Adapt a commonly used framework for evaluating MT for European languages [Callison-Burch et al., 2011]

10 bilingual annotators were hired to assess the quality of each system

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>$K_{inter}$</th>
<th>$K_{intra}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-AR</td>
<td>0.57</td>
<td>0.62</td>
</tr>
<tr>
<td>Average EN-EU</td>
<td>0.41</td>
<td>0.57</td>
</tr>
<tr>
<td>EN-CZ</td>
<td>0.40</td>
<td>0.54</td>
</tr>
</tbody>
</table>
AL-BLEU: Evaluation and Results

- Use 900 sentences extracted from the dataset: 600 dev and 300 test
- AL-BLEU correlates better with human judgments

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>0.3361</td>
<td>0.3162</td>
</tr>
<tr>
<td>AL-BLEU</td>
<td>0.3759</td>
<td>0.3521</td>
</tr>
</tbody>
</table>

\[ \tau = \left( \# \text{of concordant pairs} - \# \text{of discordant pairs} \right) \div \text{total pairs} \]
AL-BLEU: Conclusion

- We provide an annotated corpus of human judgments for evaluation of Arabic MT
- We adapt BLEU and introduce AL-BLEU
- AL-BLEU uses morphological, syntactic and lexical matching
- AL-BLEU correlates better with human judgments

http://nlp.qatar.cmu.edu/resources/AL-BLEU
Thank you for your attention
Collaborators

Prof. Kemal Oflazer  Dr. Behrang Mohit  Hanan Mohammed