Comparing RBMT and SMT systems + BLEU score

Samantha Link

6. June 2018
Thurmair (2004): Comparing Rule-based and Statistical MT Output
Compare the state-of-the-art quality of MT

A current statistical MT package and a commercial rule-based MT system were tested.

The data consisted of Translation Memory material, German to English, more than 100,000 segments.
Statistical MT

- Verbmobil project (Vogel et al. 2000)
- Input: Training corpus with 1.068 mio German and 1.128 mio English tokens representing 44.400 German and 26.600 English types.
- A test corpus (5% of the corpus) consisting of all sentences of (randomly) of 14 tokens of length containing no unknown words was analyzed (68 sentences).
SMT Evaluation

- **grammatical**: This means the sentences are syntactically correct, and convey the content.
- **understandable**: This means the sentences are incorrect but still convey the content (without reference to the source text).
- **wrong**: This means that the sentences cannot be understood without reference to the source text.

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>grammatical</td>
<td>16</td>
<td>23.5%</td>
</tr>
<tr>
<td>understandable</td>
<td>31</td>
<td>45.6%</td>
</tr>
<tr>
<td>wrong</td>
<td>21</td>
<td>30.9%</td>
</tr>
</tbody>
</table>
Rule-based MT

- Linguatec’s “Personal Translator” German-to-English
- The only action was to add some of the unknown words to the system dictionary.
- The 68 test sentences (860 words, mainly very specialised database terminology)
- About 60 were not in the system dictionary. Of those, 20 were coded, using the system’s coding tool.
RBMT Evaluation

**Question to you:** How would you evaluate the performance of the systems compared to each other?

<table>
<thead>
<tr>
<th></th>
<th>SMT</th>
<th>SMT</th>
<th>RBMT</th>
<th>RBMT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>grammatical</strong></td>
<td>16</td>
<td>23,5%</td>
<td>30</td>
<td>44,1%</td>
</tr>
<tr>
<td><strong>understandable</strong></td>
<td>31</td>
<td>45,6%</td>
<td>24</td>
<td>35,3%</td>
</tr>
<tr>
<td><strong>wrong</strong></td>
<td>21</td>
<td>30,9%</td>
<td>14</td>
<td>20,6%</td>
</tr>
</tbody>
</table>
Improvements:

”However, the question is not so much which approach is better; the more interesting question is what can be learned for the respective other approach, and how a hybrid system by which significant improvement in MT quality could be achieved should look like.”
SMT Improvements

- German verb order and Satzklammer (split verbs) phenomena (Verbs in subordinate clauses must go from German last to English second position).
- The system tends to keep the constituent order as in the source language.
- Special constructions like German conditional clauses without subjunction (translated with indicative).
- The system tends to drop pronouns.
- Morphological analysis of German noun compounds.
- Special treatment of variable and product names.
"Such mis-handlings are systematic, they are responsible for about 55% of the ‘wrong’ evaluations, and it is hard to see how they could be overcome even if the training corpus could be extended significantly, because the “normal” material always outperforms the special cases.”
RBMT Improvements

- Parse failures do not allow to identify the sentence parts; systems often use fall-back rules for those cases, but there will always be sentences which cannot be analysed properly.
- Lexical failures make up 2/3 of all errors (selecting wrongly from the dictionary).
Conclusions

"The best way to proceed seems to be to create a hybrid system based on a rule-based architecture, and enrich it by features of statistical MT."

▶ RBMT starts from a better quality baseline.
▶ SMT gets better input by preprocessed data (better segmentation, morphological decomposition, name recognition etc.).
▶ Replacing the (rather primitive) target language models by smarter linguistic-based generation components.
▶ However, grammatical reference to the source sentence is still necessary, esp. in the area of grammatical functions (subject, object etc.).
Question to you: In which ways would a RBMT system benefit from being enhanced by a statistical approach?
Introduction

BiLingual Evaluation Understudy

- Human evaluations of MT output are extensive, but also expensive and can take weeks or months to finish.
- Developers would benefit from an inexpensive automatic evaluation that is quick, language independent, and correlates highly with human evaluation.
- Central idea: The closer a machine translation is to a professional human translation, the better it is.
- A automatic evaluation system would need:
  1. a numerical ”translation closeness” metric
  2. a corpus of good quality human reference translations
Example 1:

**Candidate 1**: It is a guide to action which ensures that the military always obeys the commands of the party.

**Candidate 2**: It is to insure the troops forever hearing the activity guidebook that party direct.
The Baseline BLEU Metric

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.
More word-overlaps can be observed between the better translation Candidate 1 and the human translations.

Thus, a evaluation program can rank Candidate 1 higher than Candidate 2, based on n-gram matches.

Compare n-grams of the candidate with the n-grams of the reference translation and count the number of matches.
Modified n-gram precision

- Precision: Number of candidate translation words (unigrams) of the reference translation divided by the total number of words in the candidate translation.

- **Question to you:** Do you see any problems here?
Modified n-gram precision

- Precision: Number of candidate translation words (unigrams) of the reference translation divided by the total number of words in the candidate translation.
- Problem of overgeneration as in example 2:

Example 2.

Candidate: the the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Standard Unigram Precision = 7/7.
A reference word should be considered exhausted after a matching candidate word is identified.

▶ Count the maximum number of times a word occurs in any single reference translation.
▶ Clip the total count of each candidate word by its maximum reference count.
▶ Add these clipped counts up, and divide by the total (unclipped) number of candidate words.
Exercise: Modified n-gram precision

**Question to you:** What is the modified unigram precision of the sentences in Example 1?

**Candidate 1:** It is a guide to action which ensures that the military always obeys the commands of the party.

**Candidate 2:** It is to insure the troops forever hearing the activity guidebook that party direct.

**Reference 1:** It is a guide to action that ensures that the military will forever heed Party commands.

**Reference 2:** It is the guiding principle which guarantees the military forces always being under the command of the Party.

**Reference 3:** It is the practical guide for the army always to heed the directions of the party.
Performance of Modified n-gram precision

Figure 1: Distinguishing Human from Machine
Performance of Modified n-gram precision

Figure 2: Machine and Human Translations

![Graph showing precision vs. phrase (n-gram) length for different methods: H2, H1, S3, S2, S1.](image-url)
Combining the modified n-gram precisions

- Modified n-gram precision decays roughly exponentially with \( n \).
- A reasonable averaging scheme must take this into account: a weighted average of the logarithm of the modified precisions would do so.
- BLEU uses the average logarithm with uniform weights, which is equivalent to using the geometric mean of the modified n-gram precisions.
- It is more sensitive to longer n-grams (best correlation with monolingual human judgments using a maximum ngram order of 4, although 3-grams and 5-grams give comparable results).
N-gram precision penalizes spurious words in the candidate that do not appear in any of the reference translations.

Additionally, modified precision is penalized if a word occurs more frequently in a candidate translation than its maximum reference count.

**Candidate:** of the

**Reference 2:** It is the guiding principle which guarantees the military forces always being under the command of the Party.

**Question to you:** What is the modified unigram and bigram precision of the candidate translation here?
The trouble with recall

- Traditionally, precision has been paired with recall to overcome such length-related problems.
- However, BLEU considers multiple reference translations, each of which may use a different word choice to translate the same source word.
The trouble with recall

Example 4:

Candidate 1: I always invariably perpetually do.

Candidate 2: I always do.

Reference 1: I always do.

Reference 2: I invariably do.

Reference 3: I perpetually do.
Sentence brevity penalty

- Candidate translations longer than their references are already penalized by the modified n-gram precision measure: there is no need to penalize them again.
- Multiplicative brevity penalty factor only penalizes candidates shorter than their reference translations.
- First compute the test corpus’ effective reference length, $r$, by summing the best match lengths for each candidate sentence in the corpus.
- The brevity penalty is a decaying exponential in $r/c$, where $c$ is the total length of the candidate translation corpus.
The BLEU Evaluation

- The BLEU metric ranges from 0 to 1. Few translations will attain a score of 1 unless they are identical to a reference translation.
- On a test corpus of about 500 sentences (40 general news stories), a human translator scored 0.3468 against four references and scored 0.2571 against two references.
When comparing human evaluation with BLEU, a high correlation coefficient of 0.99 indicates that BLEU tracks human judgment well.
Criticism

”A candidate translation should be neither too long nor too short, and an evaluation metric should enforce this. To some extent, the ngram precision already accomplishes this. Ngram precision penalizes spurious words in the candidate that do not appear in any of the reference translations. Additionally, modified precision is penalized if a word occurs more frequently in a candidate translation than its maximum reference count.”

**Question to you:** What are the advantages and disadvantages of BLEU-score?
Reddy, S., D. Raghu, M. M. Khapra & S. Joshi (2017). Generating Natural Language Question-Answer Pairs from a Knowledge Graph Using a RNN Based Question Generation Model:

- BLEU does not capture the true performance of the system.
- If the trained model simply reproduces all keywords in the generated question then the unigram overlap will also be high.
- BLEU score can be unnecessarily harsh on a model even if it generates a valid paraphrase.
- They use it in combination with human evaluation.
Thank you!