A Challenge Set Approach to Evaluating Machine Translation

Pierre Isabelle
With Colin Cherry and George Foster

EMNLP 2017, Copenhagen, 9-11 September 2017
In one slide…

• We propose a Challenge Set for English to French translation
  - Hand-crafted, short, difficult sentences
  - Each exhibiting a specific linguistic issue
  - Feeding a targeted manual evaluation

• Used to evaluate phrase-based and neural systems
• Reveals strengths and weaknesses of neural MT
Motivation

- Lots of recent excitement generated by NMT
- We trained up our own English-French system using Nematus
  - We were impressed by the results!
  - Wanted to quantify and track which tricky translation issues had been resolved, and which haven’t.
Setting the stage: Phrase-based MT

• Builds a target sentence from left to right
  - Each step translates a phrase in the source, and appends it to a growing target sentence
  - Goal is to cover all source words exactly once

• Translation is very fast
• Uses big human-licensed segments
• Corpus-specific biases are baked in
Setting the stage: Neural MT

- Essentially a target language model conditioned on the source sentence
- **Encoder** transforms the source sentence into a sequence of source summaries, each focused on a particular word
- **Attention** (softly) selects the best source summary for the current time step
- **Decoder** models the next word, given target history and source context

- End-to-end training
- Rich modelling of source and target context
- Can potentially model very long distance dependencies

Credit: OpenNMT.net
Previous work: NMT successes

• Lots of recent successes for NMT
  - Many WMT 2016 wins (Bojar et al., 2016)
  - Google’s switch to NMT (Wu et al., 2016)

• Both accompanied by pairwise human evals
Previous work: Error Analysis

• Bentivogli et al. (2016): look at post-edited IWSLT data
  - Substantial improvements in lexical, morphological and word order errors

• Toral and Sanchez-Cartagena (2017) looked at WMT
  - Similar broad conclusions
  - Marked degradation in NMT as sentence length increased

• Sennrich (2016): pairwise comparisons between two NMT systems on original and corrupted references
  - Character-based model improves generalization on unseen words.
  - Introduces some grammatical errors.
The Challenge Set

<table>
<thead>
<tr>
<th>Source</th>
<th>The repeated calls from his mother should have alerted us.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>Les appels répétés de sa mère <em>auraien</em> dû nous alerter.</td>
</tr>
<tr>
<td>System</td>
<td>Les appels répétés de sa mère <em>devraient</em> nous avoir alertés.</td>
</tr>
<tr>
<td>Is the subject-verb agreement correct? (y/n)</td>
<td></td>
</tr>
</tbody>
</table>

- Each sentence hand-designed to exhibit a *structural divergence* - a linguistic structure that does not easily map across these two languages
- Label each sentence with what it is testing - evaluate translation in terms of only that specific linguistic phenomenon
- Can provide an alternate view of translation quality - designed to complement evaluation on randomly selected “found text”
The Benefits of Targeted Sentences

- No need to weigh different types of errors against each other.
- Fast to evaluate, high agreement.
- Allows for fine-grained characterization of capabilities.

<table>
<thead>
<tr>
<th>Source</th>
<th>The repeated calls from his mother should have alerted us.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>Les appels répétés de sa mère <em>auraient</em> dû nous alerter.</td>
</tr>
<tr>
<td>System</td>
<td>Les appels répétés de sa mère devraient nous avoir alertés.</td>
</tr>
<tr>
<td>Is the subject-verb agreement correct? (y/n)</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Constructing the Challenge Set

• Included:
  - Known structural divergences.
  - Weaknesses of phrase-based MT.

• Explicitly didn’t test robustness to sparse data
  - All words occur at least 100 times in our training corpus.
  - Would like to eventually ensure that the syntactic patterns we are testing occur frequently in the training data.
Morpho-syntactic divergences

• French is morphologically richer than English:
  - French: 30 verb inflections, English: 5.
  - Person, number, gender information cannot be copied over from source word: need to be recovered from context.

• Can test specific French rules:
  - “the princess, the queen, and the woman” is feminine
  - “the princess, the queen, the king and the woman” is masculine

• Can test robustness to distractors:
  - The repeated calls from his mother should have alerted us.
    ↑ plural    ↑ singular    ↑ subj agree?
Lexico-syntactic divergences

- A specific governing word has different requirements on its arguments after translation:

**English**

Send something **to** someone.

Send someone something.

**French**

Envoyer qqch à qqun.
[Send something **to** someone]

Envoyer à qqun qqch.
[Send **to** someone something]
Syntactic divergences

• Some syntactic patterns in the source simply aren’t available in the target, for example:

• French pronouns are pro-cliticized:
  - He gave it to her.
  - Il le lui a donné. [He it her gave.]

• And you can’t get away with stranding prepositions in French (something I always get away with in English)
Morphology to reveal understanding: Who is being arrogant?

• She **asked** her brother not to be arrogant.
  → Elle a **demandé** à son frère de ne pas être **arrogant**.

• She **promised** her brother not to be arrogant.
  → Elle a **promis** à son frère de ne pas être **arrogante**.
Evaluation Systems: Data

- Challenge set is 108 hand-crafted sentences:
  - At least 3 sentences per divergence.
  - All words are frequent in training corpus.

- Systems trained on LIUM shared-task subset of the WMT 2014 corpora (12.1M sentences).

- We calculate BLEU on the WMT14 test set (3K sentences) for calibration.
Evaluation Protocol

- Three bilingual evaluators judged system outputs:
  - Answered yes-no questions.
  - Judged only the phenomenon of interest - other errors were ignored.
  - Blind to system identity.
- Provided a question and an example reference.
- We used CrowdFlower to quickly build an interface
  - in-house annotators (but not the authors)
Annotator Agreement

<table>
<thead>
<tr>
<th>Component</th>
<th>% Complete Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>MorphoSyn</td>
<td>94</td>
</tr>
<tr>
<td>LexicoSyn</td>
<td>94</td>
</tr>
<tr>
<td>Syntactic</td>
<td>81</td>
</tr>
<tr>
<td>Overall</td>
<td>89</td>
</tr>
</tbody>
</table>
Evaluation Systems: Phrase-based (PBMT)

• Strong Portage phrase-based system:
  - 2 word alignments.
  - NNJM (Devlin et al., 2014).
  - Hierarchical reordering model (Galley and Manning 2008).
  - 10K sparse features (Cherry, 2013).
  - Batch-lattice MIRA tuning (Cherry and Foster, 2012).

• Two variants:
  - PBMT1: LM built only on parallel data (data equivalent to NMT)
  - PBMT2: adds LM built on 15.1M sentences of monolingual text
Evaluation Systems: In-house Neural (NMT)

- Nematus model (single layer, GRU)
- 90K source- and target-word vocabularies built with joint byte-pair encoding (Sennrich et al., 2016)
- 512-d embeddings, 1024-d states
  - 172M parameters total
- Adadelta with gradient clipping for optimization
- Decoding with AmuNMT with a beam size of 4
Evaluation Systems: Google Neural (GNMT)

- Google recently went neural (Wu et al., 2016).
- 8 layers for both encoder and decoder.
- Residual connections.
- Data is “two to three decimal orders of magnitude bigger than the WMT corpora”.
Challenge Set Performance

- **MorphoSyn**
- **LexicoSyn**
- **Syntactic**
- **Overall**

<table>
<thead>
<tr>
<th>Method</th>
<th>PBMT1</th>
<th>PBMT2</th>
<th>NMT</th>
<th>GNMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>MorphoSyn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LexicoSyn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syntactic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- PBMT1
- PBMT2
- NMT
- GNMT
Challenge Set vs BLEU

- **WMT14 BLEU**
  - PBMT1
  - PBMT2
  - NMT

- **Challenges Correct**
  - PBMT1
  - PBMT2
  - NMT
But what about our examples?

<table>
<thead>
<tr>
<th>Source</th>
<th>The repeated calls from his mother should have alerted us.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>Les appels répétés de sa mère <strong>auraient</strong> dû nous alerter.</td>
</tr>
<tr>
<td>Is the subject-verb agreement correct? (y/n)</td>
<td></td>
</tr>
<tr>
<td>PBMT</td>
<td>Les appels répétés de sa mère <strong>aurait</strong> dû nous a alertés. <strong>X</strong></td>
</tr>
<tr>
<td>NMT</td>
<td>Les appels répétés de sa mère <strong>devraient</strong> nous avoir alertés. <strong>✓</strong></td>
</tr>
<tr>
<td>GNMT</td>
<td>Les appels répétés de sa mère <strong>auraient</strong> dû nous alerter. <strong>✓</strong></td>
</tr>
</tbody>
</table>

Subject-verb agreement across distractors
NMT Strengths: Morpho-Syntactic
16% (PBMT) => 72% (NMT)

Most cases of complex S-V agreement correctly handled

- Agreement features correctly passed across distractors
  - As in previous example.

- Agreement features correctly distributed across coordinated verb phrases
  - The woman was very tall and extremely strong.
    La femme [F-S] était très grande [F-S] et très forte [F-S]

- Assign correct agreement features to most coordinated subjects
  - The cow and the hen must be fed.

- Past Participle agreement (notoriously complex rules) is mostly correct
  - John sold the car that he had won [F-P] in a lottery.
NMT Strengths: Lexico-Syntactic
42% (PBMT) => 52% (NMT) => 62% (GNMT)

• Correctly handles double object constructions:
  John told the kids a nice story. → John a raconté aux enfants une belle histoire.
  (John told to the kids a nice story.)

• Correctly discriminates overlapping subcat frames:
  Paul knows this story. → Paul connaît cette histoire.
  Paul knows this story is hard to believe. → Paul sait que cette histoire est difficile à croire.

• Better handling of infinitival → finite complements:
  She wanted her mother to let her go. → Elle voulait que sa mère la laisse partir.
  (that her mother let [SUBJ-PRES] her go)
NMT Strengths: Purely Syntactic
33% (PBMT) => 40% (NMT) => 75% (GNMT)

• Yes/No question syntax handled correctly

Have the kids ever watched that movie? → Les enfants ont-ils déjà regardé ce film? [The kids have-they ever watched that movie?]

• Pronouns are mostly pro-cliticized correctly (i.e. attached to the left of the main verb, reflecting the person/number/case of the complement)

He gave it to her. → Il le lui a donné. [He it her gave.]

He did not talk to them very often. → Il ne leur a pas parlé très souvent. [He not them talked very often.]
GNMT Additional Strengths (purely syntactic)

• English tag questions
  
  She was perfect tonight, wasn’t she?
  
  → Elle était parfaite ce soir, n’est-ce pas? [... is this not?]

• « Inalienable possession » construction (most cases)

  I brushed my teeth.
  
  → Je me suis brossé les dents. [I brushed the teeth to myself].

• Stranded (or dangling) prepositions
## Zoom in on dangling prepositions

<table>
<thead>
<tr>
<th>Source</th>
<th>The city <strong>that</strong> he is arriving <strong>from</strong> is dangerous.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>La ville <em>d’où [from where]</em> il arrive est dangereuse.</td>
</tr>
<tr>
<td>PBMT</td>
<td>La ville <em>qu’ [that]</em> il est arrivé <em>de [from]</em> est dangereuse.</td>
</tr>
<tr>
<td>NMT</td>
<td>La ville <em>qu’ [that]</em> il est en train d’arriver est dangereuse.</td>
</tr>
<tr>
<td>GNMT</td>
<td>La ville <em>d’où [from where]</em> il vient est dangereuse.</td>
</tr>
</tbody>
</table>

The GNMT translation correctly places the preposition "from" in the French sentence, while the other systems incorrectly place it as "de" or "qu’".
NMT Weaknesses

• Big advantage of the challenge set is it can help pinpoint specific weaknesses

• Sure NMT is “strong for morphology” - but are there morphological cases it still can’t get?

• One weakness you won’t see in this survey:
  • degradation with sentence length
  • a blind spot for our strategy because we use short sentences
NMT Weaknesses: Agreement through control verbs

<table>
<thead>
<tr>
<th>Source</th>
<th>She promised her brother not to be arrogant.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>Elle a promis à son frère de ne pas être <em>arrogante</em>.</td>
</tr>
<tr>
<td>Is the subject-verb agreement correct? (y/n)</td>
<td></td>
</tr>
<tr>
<td>PBMT</td>
<td>Elle a promis son frère à ne pas être <em>arrogant</em>. x</td>
</tr>
<tr>
<td>NMT</td>
<td>Elle a promis à son frère de ne pas être <em>arrogant</em>. x</td>
</tr>
<tr>
<td>GNMT</td>
<td>Elle a promis à son frère de ne pas être <em>arrogant</em>. x</td>
</tr>
</tbody>
</table>
NMT Weaknesses: Lexically triggered exceptions

• French is a SVO language, like English, but “to miss” triggers a rare subject-object inversion:
  - Mary misses Jim.
  - Jim manque à Mary.

• You cannot “swim across something” in French, instead, you “cross something by swimming”
  - Same for other movement words (i.e.: run)

• What do these have in common?
  - A large change triggered by a small set of words
### NMT Weaknesses: Idioms

<table>
<thead>
<tr>
<th>Source</th>
<th>You are putting the cart before the horse.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>Vous mettez la charrue [plow] devant les bœufs [oxen].</td>
</tr>
</tbody>
</table>

Is the English idiomatic expression correctly rendered with a suitable French idiomatic expression?

| PBMT   | Vous pouvez mettre la charrue avant les bœufs. | ✔️ |
|--------|---------------------------------------------|
| NMT    | Vous mettez la charrue [plow] avant le cheval [horse]. | ✗ |
| GNMT   | Vous mettez le chariot [cart] devant le cheval [horse]. | ✗ |
NMT Weaknesses: Incomplete Generalizations

• Several cases where NMT appears to have captured a linguistic rule, but fails to generalize in unexpected ways, i.e.:

• The French subjunctive mood is triggered lexically:
  - NMT is good at this in general.
  - But some common triggers (“provided that”) seem not to have been captured.

• French makes some implicit noun-phrase relations explicit:
  - Knows that a water filter [\(\rightarrow\) filtre à eau] is for filtering water
  - Knows that a metal filter [\(\rightarrow\) filtre en métal] is made out of metal
  - But thinks a paper filter [\(\rightarrow\) filtre à papier] is for filtering paper!

• Would like to develop methods to find how these errors relate to characteristics of the NMT engine’s training data.
Conclusions

• Presented a challenge set methodology for MT evaluation and error analysis:
  - Provides insight into how NMT improves over PBMT
  - Also into where NMT needs to improve

• Not intended to replace automatic or manual evaluation on found text, but *supplement* it:
  - It’s not enough to get a good challenge set score.

• Full dataset is available in human and machine-readable formats, along with system outputs and human judgments.
Future Work

• Use the challenge set to evaluate and characterize any differences that come with new architectures (i.e.: fairseq):
  - the challenge set should grow as MT evolves

• Find instances of challenge set phenomena in our training text

• Automate the construction of the challenge set
  - how to automatically detect a structural divergence?

• Remove or expedite the human evaluation process

• Improve MT performance on the remaining difficult cases
  - specially designed curriculum to address incomplete generalizations
  - architecture changes to aid capturing failed generalizations
Epilogue:

DEEPL Machine Translation
Vs our Challenge Set

Pierre Isabelle
Medium post, 20 Sept. 2017
Addendum: the Buzz about DEEPL

• A buzz recently emerged about a system (impressively) said to be significantly better than GNMT: DEEPL (based on Linguee corpus).
• Great opportunity for us to check the power of our challenge set approach in assessing such a buzz.
• Overall success rates:

<table>
<thead>
<tr>
<th>System</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBMT-1</td>
<td>31%</td>
</tr>
<tr>
<td>PBMT-2</td>
<td>32%</td>
</tr>
<tr>
<td>NMT</td>
<td>53%</td>
</tr>
<tr>
<td>GNMT</td>
<td>68%</td>
</tr>
<tr>
<td>DEEPL</td>
<td>84%</td>
</tr>
</tbody>
</table>
## Success Rates on the Challenge Set

<table>
<thead>
<tr>
<th>System</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBMT-1</td>
<td>31%</td>
</tr>
<tr>
<td>PBMT2</td>
<td>32%</td>
</tr>
<tr>
<td>NMT</td>
<td>51%</td>
</tr>
<tr>
<td>GNMT</td>
<td>68%</td>
</tr>
<tr>
<td>DEEPL</td>
<td>84%</td>
</tr>
</tbody>
</table>

DEEPL’s error reduction relative to GNMT: $16/32 = 50\%$!
About DEEPL’s Performance

• Stronger on many of GNMT weak points:
  ➢ Fewer incomplete generalizations.
  ➢ Somewhat better with subject control, argument switch, manner-of-movement, etc.
  ➢ But only marginally better with idioms!

• Source of gains is still uncertain:
  ➢ Probably not structure of NN model
  ➢ Probably not training data size
  ➢ Perhaps training data quality?

• So good that… our CS may already have become too easy!
• But we know how to make it harder…
The Hardest Problems for MT

• Common sense reasoning may be needed any time

• Example 1: Pronoun reference and translation:

The city councillors refused the women a demonstration permit because { they (→ ils) feared they (→ elles) advocated } violence.

• Example 2: Word sense disambiguation

Il a payé ses études en vendant de l’assurance. → … by selling insurance.

Il a réglé la note en finissant son café. → … while finishing his coffee.
Questions?