Gap-filling as a method to evaluate the usefulness of raw machine translation

Mikel L. Forcada

Departament de Llenguatges i Sistemes Informàtics
Universitat d’Alacant
E-03071 Alacant (Spain)

SMT Group, Edinburgh University, 7th September 2016
1. Most machine translation is used raw
2. What’s special about MT to be used raw?
3. Evaluation of MT to be used raw
4. An alternative: gap-filling evaluation
5. A survey of existing work
   - Unhammer and Trosterud (2012)
   - O’Regan and Forcada (2013)
   - Ageeva et al. (2015)
   - Jordán et al. (2016) (yet unpublished)
6. Critique
7. My work in Edinburgh
Outline

1. Most machine translation is used raw
2. What’s special about MT to be used raw?
3. Evaluation of MT to be used raw
4. An alternative: gap-filling evaluation
5. A survey of existing work
   - Unhammer and Trosterud (2012)
   - O’Regan and Forcada (2013)
   - Ageeva et al. (2015)
   - Jordán et al. (2016) (yet unpublished)
6. Critique
7. My work in Edinburgh
F.J. Och, Google, 2012:

In a given day we translate roughly as much text as you’d find in 1 million books. To put it in another way: what all the professional human translators in the world translate in a year, our system translates in roughly a single day.
Most MT is used raw /2

Online MT is readily available almost to anyone:

- Seamlessly incorporated into online applications (for example browsers)
- Integrated in main digital social media (for example, Twitter)

Raw MT may sometimes be the only option:

- user-generated content
- when quickly browsing to select what to read
Raw MT is sometimes provided instead of professional translations. Some examples:

- Microsoft offers machine-translated version of support articles in their Knowledge Base for more than one decade
- TripAdvisor uses raw machine translation into Russian provided by ProMT.
- Amazon’s portal in Spain shows descriptions of products in raw machine-translated Spanish
Outline

1. Most machine translation is used raw
2. What’s special about MT to be used raw?
3. Evaluation of MT to be used raw
4. An alternative: gap-filling evaluation
5. A survey of existing work
   - Unhammer and Trosterud (2012)
   - O’Regan and Forcada (2013)
   - Ageeva et al. (2015)
   - Jordán et al. (2016) (yet unpublished)
6. Critique
7. My work in Edinburgh
Raw MT is clearly very different from natively-produced or professionally-translated text.

Surprisingly, very little research has been done on how readers make sense of MT output:

- Is it similar to understanding a non-native ("broken") version of their language?
- Is raw MT similar to a contact language (Pidgin or Creole) based on the target language? (old idea: Masterman 1967)
- Is it similar to understanding a closely-related foreign language?
- Do readers *adapt* similarly to how they do with non-natives? Do they *learn* and become better at the task?
What’s special about MT to be used raw?

To improve MT for **its main usage**, we need:

- Fast methods to measure its usefulness.
- Knowledge about how it is actually understood. Can some “nativeness” be thrown away without damage?
  - Definite articles (imagine a Russian speaking English)
  - Phrase order (“We the Yoda easily understand”)
  - Gender and number agreement in noun phrases (English already did away with it!)
Outline

1. Most machine translation is used raw
2. What’s special about MT to be used raw?
3. Evaluation of MT to be used raw
4. An alternative: gap-filling evaluation
5. A survey of existing work
   - Unhammer and Trosterud (2012)
   - O’Regan and Forcada (2013)
   - Ageeva et al. (2015)
   - Jordán et al. (2016) (yet unpublished)
6. Critique
7. My work in Edinburgh
Evaluating MT for *gisting* or *assimilation* is not easy.

- Extrinsic evaluation may involve assessing actual task completion through the use of machine-translated instructions (Castilho et al. 2014; Doherty and O’Brien 2013).
- But this is expensive: are there alternatives to actual task-oriented evaluation that correlate well with it?
Alternatives to task-oriented evaluation:

- Standard approaches use a costly “reading comprehension” approach with carefully-crafted TL questions (Jones et al. 2007).
- Alternative methods based on blind post-editing followed by human assessment of adequacy by bilinguals are also expensive (WMT 2009, 2010; Ginestí-Rosell et al. 2009).
- We want a less expensive way to evaluate how much MT improves understanding of foreign text.
Outline

1. Most machine translation is used raw
2. What’s special about MT to be used raw?
3. Evaluation of MT to be used raw
4. An alternative: gap-filling evaluation
5. A survey of existing work
   - Unhammer and Trosterud (2012)
   - O’Regan and Forcada (2013)
   - Ageeva et al. (2015)
   - Jordán et al. (2016) (yet unpublished)
6. Critique
7. My work in Edinburgh
A novel cloze test (closure test) strategy, starting with a parallel corpus

- Cloze tests had so far been performed on raw MT output, not on reference sentences (Somers and Wild 2000).

- Gap-filling technique devised by Francis Tyers and Mikel Forcada and informally spread across the Apertium community.
The procedure:

- Create **holes** or **gaps** in the reference target-language (TL) sentences by **blanking out** part of the words
  - Blanked-out words marked by a placeholder, e.g. ######
  - One may blank out non-stopwords only, or specific parts of speech such as nouns.
  - One may blank out every $n$-th word, or randomly with some probability.
The procedure:

- Ask non-SL-speaking subjects to fill the gaps in randomly chosen TL sentences in different hinting situations and measure the success rate:
  - Without any hint whatsoever (guesswork: lower bound)
  - Showing the SL sentence (expected to help little)
  - Showing the TL sentence produced by MT (this is what we want to measure!)
  - Showing an alternative TL reference translation (paraphrase: upper bound)
- One or more of these hints may be shown.
An example: **no hint**

| Problem sentence: | The ###### of the ###### Declaration have addressed in ###### ###### the demands of the international ###### which the ###### Declaration represents. |
An example: **Source-language hint**

<table>
<thead>
<tr>
<th>Basque (source language) hint:</th>
<th>Bruselako Adierazpenaren sinatzaileek argi eta garbi zuzendu dute Adierazpen horrek ordezkatzen duen nazioarteko komunitatearen eskaera.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem sentence:</td>
<td>[sm]@@143: The ###### of the ###### Declaration have addressed in ###### ###### the demands of the international ###### which the ###### Declaration represents.</td>
</tr>
</tbody>
</table>
An example: **MT hint**

<table>
<thead>
<tr>
<th>Machine translation hint:</th>
<th>the signatories of the Statement of Brussels clear and clean they have addressed this Statement he of the international community that substitutes the request.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem sentence:</td>
<td>[sm]@@143: The ###### of the ###### Declaration have addressed in ###### ###### the demands of the international ###### which the ###### Declaration represents.</td>
</tr>
</tbody>
</table>
An example: **Both hints**

| Basque (source language) hint: | Bruselako Adierazpenaren sinatzaileek argi eta garbi zuzendu dute Adierazpen horrek ordezkatzen duen nazioarteko komunitatearen eskaera. |
| Machine translation hint: | the signatories of the Statement of Brussels clear and clean they have addressed this Statement he of the international community that substitutes the request. |
| Problem sentence: | [sm]@@143: The ###### of the ###### Declaration have addressed in ###### ###### the demands of the international ###### which the ###### Declaration represents. |
An example: **How did we do?**

<table>
<thead>
<tr>
<th>Basque (source language) hint:</th>
<th>Bruselako Adierazpenaren sinatzaileek argi eta garbi zuzendu dute Adierazpen horrek ordezkatzen duen nazioarteko komunitateareneko eskaera.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine translation hint:</td>
<td>the signatories of the Statement of Brussels clear and clean they have addressed this Statement he of the international community that substitutes the request.</td>
</tr>
<tr>
<td>Problem sentence:</td>
<td>[sm]@@143: The ###### of the ###### Declaration have addressed in ###### ###### the demands of the international community which the ###### Declaration represents.</td>
</tr>
<tr>
<td>Reference sentence:</td>
<td>The <strong>endorsers</strong> of the <strong>Brussels</strong> Declaration have addressed in <strong>unequivocal terms</strong> the demands of the international <strong>community</strong> which the <strong>Brussels</strong> Declaration represents.</td>
</tr>
</tbody>
</table>
“Synonyms” may (optionally) be allowed:

<table>
<thead>
<tr>
<th>endorsers</th>
<th>signatories</th>
</tr>
</thead>
<tbody>
<tr>
<td>unequivocal</td>
<td>clear</td>
</tr>
<tr>
<td>mesures</td>
<td>measures</td>
</tr>
<tr>
<td>mandate</td>
<td>Mandate</td>
</tr>
<tr>
<td>likewise</td>
<td>also</td>
</tr>
<tr>
<td>legalization</td>
<td>legalisation</td>
</tr>
<tr>
<td>lawful</td>
<td>legitimate</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Another possibility is to show a drop-down menu of possible solutions for each placeholder.
Outline

1. Most machine translation is used raw
2. What’s special about MT to be used raw?
3. Evaluation of MT to be used raw
4. An alternative: gap-filling evaluation
5. A survey of existing work
   - Unhammer and Trosterud (2012)
   - O’Regan and Forcada (2013)
   - Ageeva et al. (2015)
   - Jordán et al. (2016) (yet unpublished)
6. Critique
7. My work in Edinburgh
Set out to evaluate the rule-based Apertium system from North Sámi (an Uralic language\(^1\)) to Norwegian Bokmål (Indoeuropean Germanic)
  (designed for gisting purposes)

Ten consecutive sentences, 5 to 15 words long.

At least two nouns removed.

Randomly ordered drop-down list containing the correct noun and many others, as well as a “none of the above” option.

Ten informants.

MT always shown as a hint.

---

\(^1\) Disapproved name: *Lapp*
Original: Lahttu, gean stáhta lea nammadan, lea lávdegotti jodiheaddji
(‘The member, who the Government has instituted, is the committee’s leader’)
Translated: Medlemmet, som staten har oppnevnt, det er komiteens leder
(‘The member, who the Government has instituted, that is the leader of the committee’)

Pick the right alternative: Medlemmet oppnevnt av
☐ GRUNNEN ☐ STATEN ☐ KOMMUNEN ☐ LEIDEREN ☐ KOMITEEN ☐ HAUSTAREN
☐ (INGEN SOM PASSER)
er leder for
☐ SJØEN ☐ KOMITEEN ☐ VURDERINGEN ☐ DOMMEN ☐ GANGEN ☐ UTGANGEN
☐ (INGEN SOM PASSER)
(‘The member instituted by the
go of grounds ☐ government ☐ county ☐ leader ☐ committee
☐ harvester ☐ (none of the above)
is the leader of the
☐ sea ☐ committee ☐ assessment ☐ verdict ☐ time ☐ closing ☐ (none of the above)’

M.L. Forcada (Universitat d’Alacant)
Informants successfully filled about 71% of the noun gaps.
Set out to evaluate the rule-based Apertium Basque (SOV, agglutinative, language isolate) to English (Indoeuropean, little inflection, SVO), rev. 39606.

(also designed for gisting evaluation)

20% of content (non-stopword) words removed at random.

Each informant solved 32 sentences from the same source.

Four hinting situations: no hint, source only, MT only, source and MT (roughly 8 problems of each kind).

Twenty-three informants with good English command and no Basque at all.

Post-hoc synonym list (86 entries) conservatively built after inspecting results.
Success is high with no hints (repetitive, predictable text)
SL hint not too useful (proper nouns and cognates?)
Success rate improves clearly with MT hint
Having both hints seems to hurt
Synonyms do not change general trend

<table>
<thead>
<tr>
<th>HINT LEVEL</th>
<th># OF 1-WORD GAPS</th>
<th>SUCCESS RATE (EXACT)</th>
<th>SUCCESS RATE (SYNONYMS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No hint</td>
<td>575</td>
<td>26% (sd 13%)</td>
<td>30% (sd 14%)</td>
</tr>
<tr>
<td>SL hint</td>
<td>543</td>
<td>29% (sd 12%)</td>
<td>34% (sd 14%)</td>
</tr>
<tr>
<td>MT hint</td>
<td>597</td>
<td>48% (sd 13%)</td>
<td>54% (sd 13%)</td>
</tr>
<tr>
<td>Both hints</td>
<td>589</td>
<td>43% (sd 13%)</td>
<td>51% (sd 14%)</td>
</tr>
</tbody>
</table>
A survey of existing work

Ageeva et al. (2015)

Ageeva, Forcada, Tyers and Pérez-Ortiz (2015)

- Built an extension to Christian Federmann’s Appraise.
- Set out to evaluate two rule-based three Apertium systems, all designed for gisting evaluation:
  - Basque (SOV, agglutinative, isolate) to Spanish (Indoeuropean Romance, SVO)
  - Tatar (SOV, agglutinative, Turkic) to Russian (Indoeuropean Slavic, SVO).
- 10%, 20% or 30% of content (non-stopword) words removed.
- Each informant solved 36 sentences from the same source, all longer than 10 words.
As O’Regan and Forcada (2013), four hinting situations: no hint, source only, MT only, source and MT (9 problems of each kind).

Unspecified number of informants with good target-language command and no source-language knowledge at all.

Post-hoc synonym list conservatively built after inspecting repeated results.

Interannotator agreement studied (Krippendorff’s $\alpha$).

Three different genres studied for Tatar–Russian.
Machine translation seems to help in general (statistical significance test performed).

Results are not too conclusive as regards variation with gap density.
Submitted to AMTA 2016, did not make it.
Set out to investigate a series of research questions relating the comprehension of raw MT:
- Is raw MT into $L_1$ similar to text in a language $L'$ closely related to $L_1$?
- How useful is raw MT into $L'$ for $L_1$ readers?
- Is non-native $L_1$ text similar to raw $L_1$ output?
- Which MT system produces the most useful $L_1$ text?

I will not cover them all here.

Source languages: English, French; $L_1 = $ Spanish; $L' = $ Italian.
Twenty-four short excerpts containing whole sentences, 65–75 words long.

One every 5 words removed (any kind of words).

Four genres.

Three MT systems: Systran, Google and Apertium.

Student informants (77 native Spanish speakers), English and French lower than CEFR B1.

All experiments with just MT as a hint.

Synonyms allowed.
Main results:

- Intelligibility of raw MT $L_1$ similar or better than $L'$ (except for social sciences genre).

- *Non-native* text: raw MT $L_1$ is better than *anglo*-$L_1$, but worse than *franco*-$L_1$

- In general terms, Apertium translates English better than Google or Systran (except social sciences genre).

- In general terms, Systran and Google translate French better than Apertium.
Critique

Outline

1. Most machine translation is used raw
2. What’s special about MT to be used raw?
3. Evaluation of MT to be used raw
4. An alternative: gap-filling evaluation
5. A survey of existing work
   - Unhammer and Trosterud (2012)
   - O’Regan and Forcada (2013)
   - Ageeva et al. (2015)
   - Jordán et al. (2016) (yet unpublished)
6. Critique
7. My work in Edinburgh

M.L. Forcada (Universitat d’Alacant)
Critique

- All of this work is very preliminary and needs to be completed.
  - Alternative reference translations have never been used as hints (optimal situation)
  - A more thorough study of how the success rate deteriorates with gap density for each hinting situation is needed.
  - There was no limit of time.

- It remains to be seen how gap-filling success rates correlate with the results of a task-oriented evaluation.
  - Subjects that could be very good at *playing* the gap-filling *game* may not be that good at completing a specific task using raw MT.

- Only Jordán et al. (2016) have attempted a comparative evaluation of different MT systems:
  - How would this ranking correlate with other rankings?
Outline

1. Most machine translation is used raw
2. What’s special about MT to be used raw?
3. Evaluation of MT to be used raw
4. An alternative: gap-filling evaluation
5. A survey of existing work
   - Unhammer and Trosterud (2012)
   - O’Regan and Forcada (2013)
   - Ageeva et al. (2015)
   - Jordán et al. (2016) (yet unpublished)
6. Critique
7. My work in Edinburgh
While on sabbatical from the Universitat d’Alacant I will visit the SMT group in Edinburgh.

I plan to visit on a weekly basis.

My work will be related to the *Health in my Language* (HimL) H2020 project.
Study the correlation between gap-filling success rates for different gap densities, gap generation strategies, and real measurements of raw MT usefulness

For instance, generating a PICO (patient/population – intervention/indicator – comparison/control – outcome) summary from the MT-ed description of a medical intervention

2 Study the correlation between gap-filling success rates and other subjective measures of MT quality (classifications, rankings)

3 Combine gap-filling success rates into a single measure that optimally correlates.
4. Study the correlation between gap-filling success rates for different gap densities, gap generation strategies, and real measurements of raw MT usefulness.

5. Study the possibility of organizing a gap-filling human evaluation for WMT 2017.

6. Study the impact of actually dealing with morphology in HimL languages (Romanian, Polish, Czech, German).

7. Exploring new automatic measures that track task-oriented quality, either directly (expensive) or through gap-filling (cheaper) proxy.

8. If successful, explore tuning to these measures.
License

This work may be distributed under the terms of
- the Creative Commons Attribution–Share Alike license:
  http://creativecommons.org/licenses/by-sa/3.0/
- the GNU GPL v. 3.0 License:
  http://www.gnu.org/licenses/gpl.html

Dual license! E-mail me to get the sources: mlf@ua.es