Improving SMT by learning translation direction

Cyril Goutte, David Kurokawa, Pierre Isabelle

Interactive Language Technologies group
Institute for Information Technology
National Research Council
Motivation

We address two questions:

1. Is there a difference between original and (human-) translated text and can we detect it reliably?

2. If so, can we use that to improve Machine Translation quality?
Motivation

We address two questions:

1. Is there a difference between original and (human-) translated text and can we detect it reliably?

2. If so, can we use that to improve Machine Translation quality?

Our answers:

1. Yes: on the Canadian Hansard, we get 90+% accuracy.

2. Yes: on French-English, we obtain up to 0.6 BLEU point increase.
Problem setting

Translations often have a “feel” of the original language: *Translationese*.

If *translationese* is real, it may be possible to detect it!

Earlier studies:

- Baroni&Bernardini (2006): detect original vs. translation is a monolingual Italian corpus, with accuracy up to 87%.


Both show that various aspects of *translationese* are detectable.

We experiment on a large bilingual corpus (Hansard) and investigate how detecting translation direction may impact Machine Translation quality.
Index

1 Motivation and setting ▷ 1

○ 2 Data ▷ 4

3 Detecting Translation Direction ▷ 8

4 Exploiting Translation Direction in SMT ▷ 14

5 Discussion ▷ 20
Data: The Hansard corpus

Bilingual (En-Fr) transcripts of the sessions of the Canadian parliament.


1. Tagged with information on original language (French or English).
3. Large amount of data: 4.5M sentences, 165M words.

<table>
<thead>
<tr>
<th></th>
<th>fo</th>
<th>eo</th>
<th>mx</th>
</tr>
</thead>
<tbody>
<tr>
<td>words (fr)</td>
<td>14,648K</td>
<td>72,054K</td>
<td>86,702K</td>
</tr>
<tr>
<td>words (en)</td>
<td>13,002K</td>
<td>64,899K</td>
<td>77,901K</td>
</tr>
<tr>
<td>sentences</td>
<td>902,349</td>
<td>3,668,389</td>
<td>4,570,738</td>
</tr>
<tr>
<td>blocks</td>
<td>40,538</td>
<td>42,750</td>
<td>83,288</td>
</tr>
</tbody>
</table>
Data: The Hansard corpus (II)

Corpus issues:

▶ Slightly **inconsistent** tagging, eg both sides claim to be original: puts overall tagging reliability into question.

▶ **Missing** text/alignment, eg valid English but no translation: seems to be a retrieval issue.

▶ **Imbalance** at the word/sentence level: 80% originally English.

▶ There may be lexical/contextual **hints**: Quebec MPs tend to speak French, western Canada MPs almost only anglophones.
Corpus (pre)processing

- Tokenized (NRC in-house tokenizer)
- Lowercased
- Sentence-aligned (NRC implementation of Gale&Church, 1991)

We consider two levels of granularity:
- Sentence-level: individual sentences;
- Block-level: maximal consecutive sequence with same original language.

Block-level is balanced, sentence-level is imbalanced 4:1 (eo:fo).

Tagged using freely available “Tree Tagger” (Schmid, 1994).

⇒ 4 representations: 1) word, 2) lemma, 3) POS and 4) mixed n-grams.

“Mixed”: POS for content words, surface form for grammatical words.
Index

1 Motivation and setting ▷ 1

2 Data ▷ 4

3 Detecting Translation Direction ▷ 8

4 Exploiting Translation Direction in SMT ▷ 14

5 Discussion ▷ 20
Detecting translation direction

Support Vector Machines trained with T. Joachims’ SVM-Perf.

Test various conditions:

1. Block-level (83K examples) or sentence-level (1.8M examples, balanced).
2. Features: word, lemma, POS, mixed... n-gram frequencies.
3. N-gram length: 1...3 for word/lemma, 1...5 for POS/mixed.
4. Monolingual (English or French) or bilingual text.

Sentence-level: test fewer feature/n-gram combinations (because of computational cost).

All results obtained from 10-fold cross-validation.

Results reported in $F$-score ($\approx$ accuracy in this case).
Block-level Performance

Detection performance (en)

Similar perf. on French, +1-2% for bilingual, same general general shape.

tf-idf: small but consistent improvement.

Optimal:
word/lemma bigram, POS/mixed trigram.

Word bigram: $F = 90\%$
Mixed trigram: $F = 86\%$. 
Influence of block length

Large range in block length (3-73887 words!).

Up to 99% accuracy for large blocks.

Much better than random for short blocks.

word > lemma > mixed
Sentence-level Performance

1.8M examples (balanced)

Some missing conditions (computational cost)

$F = 77\%$
Analysis of

Most important bigrams in English (eo= original, fo=translation).

Most important=relatively more frequent.

“A couple of”: no equivalent in French

Canadian alliance, CPC, NDP: mostly western, mostly anglophone parties
BQ (Bloc Quebecois): French-speaking

French translation overuses articles, prepositions (because French does), and “Mr. Speaker”!

<table>
<thead>
<tr>
<th>eo</th>
<th>fo</th>
</tr>
</thead>
<tbody>
<tr>
<td>couple_of</td>
<td>of_the</td>
</tr>
<tr>
<td>alliance_)</td>
<td>mr_</td>
</tr>
<tr>
<td>a_couple</td>
<td>_the</td>
</tr>
<tr>
<td>do_that</td>
<td>in_the</td>
</tr>
<tr>
<td>,_canadian</td>
<td>to_the</td>
</tr>
<tr>
<td>the_record</td>
<td>_i</td>
</tr>
<tr>
<td>forward_to</td>
<td>_the</td>
</tr>
<tr>
<td>,_cpc</td>
<td>)_:</td>
</tr>
<tr>
<td>cpc_)</td>
<td>speaker_,</td>
</tr>
<tr>
<td>of_us</td>
<td>_i</td>
</tr>
<tr>
<td>this_country</td>
<td>_mr</td>
</tr>
<tr>
<td>this_particular</td>
<td>_and</td>
</tr>
<tr>
<td>many_of</td>
<td>_speaker</td>
</tr>
<tr>
<td>canadian_alliance</td>
<td>bq_)</td>
</tr>
<tr>
<td>across_the</td>
<td>,_bq</td>
</tr>
<tr>
<td>out_there</td>
<td>hon_</td>
</tr>
<tr>
<td>the_things</td>
<td>that_the</td>
</tr>
<tr>
<td>for_that</td>
<td>on_the</td>
</tr>
</tbody>
</table>
Index

1 Motivation and setting ▷ 1

2 Data ▷ 4

3 Detecting Translation Direction ▷ 8

○ 4 Exploiting Translation Direction in SMT ▷ 14

5 Discussion ▷ 20
Impact on Statistical Machine Translation

Typical SMT system training:

- Gather as much English-French aligned sentences as possible.
- Preprocess + split data
- Estimate parameters in either direction (en→fr and fr→en)
- Original translation direction is not considered at all!

⇒ We use French originals and English translations to train an en→fr system (”reverse” translation??)

We know SMT is very sensitive to genre/topic…

Does difference between original and translation matter? If so, by how much?
Impact on Statistical Machine Translation

We analyze the impact of translation direction on MT by investigating:

1. Do we get better performance by sending original text to MT system trained only on original text?
Impact on Statistical Machine Translation

We analyze the impact of translation direction on MT by investigating:

1. Do we get better performance by sending original text to MT system trained only on original text?
2. Detecting translation direction and sending text to the “right” MT system.
Impact of Original Language

System trained on eo, fo, or mx, tested on eo/fo part of test set, or all (mx).

<table>
<thead>
<tr>
<th>Train</th>
<th>mx test set</th>
<th>fo test set</th>
<th>eo test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fr $\rightarrow$ en</td>
<td>en $\rightarrow$ fr</td>
<td>fr $\rightarrow$ en</td>
</tr>
<tr>
<td>mx</td>
<td>36.2</td>
<td>37.1</td>
<td>36.1</td>
</tr>
<tr>
<td>fo</td>
<td>31.2</td>
<td>30.8</td>
<td>36.2</td>
</tr>
<tr>
<td>eo</td>
<td>36.6</td>
<td>37.8</td>
<td>33.7</td>
</tr>
</tbody>
</table>

eo system does (much) better on eo test, with 80% of training data.

eo system also does better on mx data (test is 88% eo data vs. 80% in train).

fo system does much worse on mx and eo data, but about the same as mx on the fo data, with only 20% of the training data!

⇒ Idea: detect source language using classifier, then use the right MT system (“Mixture of Experts”)

Cyril Goutte
Impact of Automatic Detection

Top part is more or less identical to previous table.

<table>
<thead>
<tr>
<th></th>
<th>Full test set</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fr→en</td>
<td>en→fr</td>
</tr>
<tr>
<td>mx</td>
<td>36.86</td>
<td>37.78</td>
</tr>
<tr>
<td>fo</td>
<td>32.00</td>
<td>31.85</td>
</tr>
<tr>
<td>eo</td>
<td>37.20</td>
<td>38.23</td>
</tr>
<tr>
<td>SVM</td>
<td>37.44</td>
<td>38.35</td>
</tr>
<tr>
<td>ref</td>
<td>37.46</td>
<td>38.35</td>
</tr>
</tbody>
</table>

ref: using reference source language information, gain a consistent ≈ 0.6 BLEU points.

SVM: using SVM prediction, gain is similar.

Smaller gain over the eo system (due to having 88% eo data in test set).

⇒ Detecting original vs. translation provides a small-ish but consistent improvement in translation performance.

⇒ not worth looking for better classifier (for that task).

Other uses of translation direction detection?
Index

1 Motivation and setting  ▶ 1

2 Data ▶ 4

3 Detecting Translation Direction ▶ 8

4 Exploiting Translation Direction in SMT ▶ 14

○ 5 Discussion ▶ 20
Discussion

How general are these results? Will it generalize to:

1. Detection on other English-French data?
2. Training a classifier on another corpus?
3. Another language pair?
4. Other settings: source vs. translations from different languages.

Mixture of experts: could use additional input-specific information.

- Mother tongue?
- Gender?
To Conclude...

Can we tell the difference between an original and translated document?
→ Yes.

To what level of accuracy?
→ Up to 90+% accuracy on blocks, 77% on single sentences.

Is translation direction useful for machine translation?
→ Yes!

Is the classification performance sufficient?
→ Indistinguishable from reference labels...
Index

1 Motivation and setting ▷ 1
2 Data ▷ 4
3 Detecting Translation Direction ▷ 8
4 Exploiting Translation Direction in SMT ▷ 14
5 Discussion ▷ 20