National Research Council Canada

Conseil national de recherches Canada

NRC · CNRC Institute for Information Technology

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%.
- van Halteren (2008) : detect source language in multi-parallel corpus and identify source language markers.

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

Data: The Hansard corpus

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

Most of 35th to 39th parliaments, covering 1996-2007.

1. Tagged with information on original language (French or English).
2. High quality translation: Reference material in Canada.
3. Large amount of data: 4.5M sentences, 165M words.

	fo	eo	mx
words (fr)	14,648K	72,054K	86,702K
words (en)	13,002K	64,899K	77,901K
sentences	902,349	3,668,389	4,570,738
blocks	40,538	42,750	83,288

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).

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

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

Index

Detecting translation direction

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

Test various conditions:

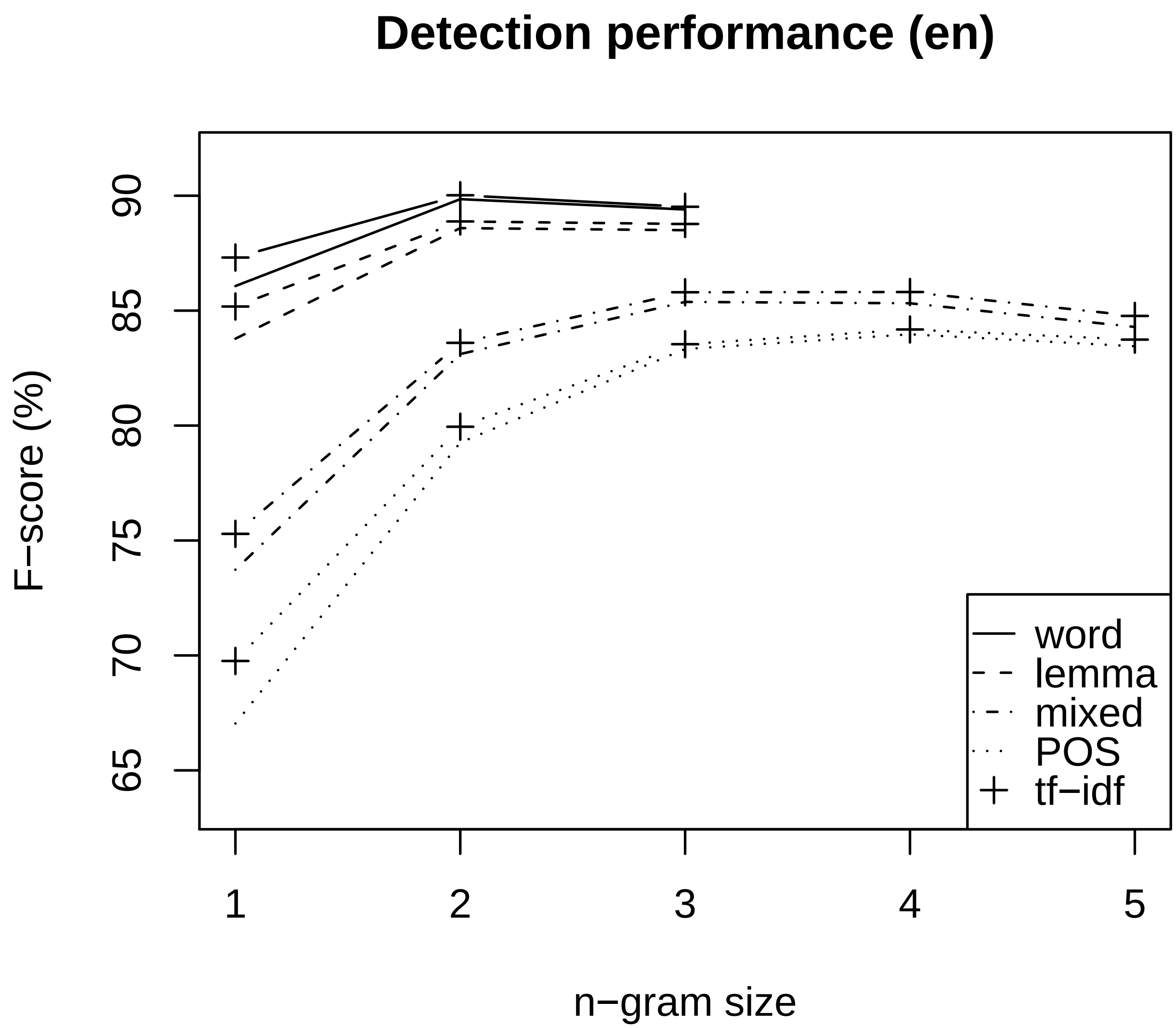
1. Block-level (83K examples) or sentence-level (1.8M examples, balanced).
2. Features: word, lemma, POS, mixed. . . n-gram frequencies.
3. N-gram length: 1. . . 3 for word/lemma, 1. . . 5 for POS/mixed.
4. Monolingual (English or French) or bilingual text.

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



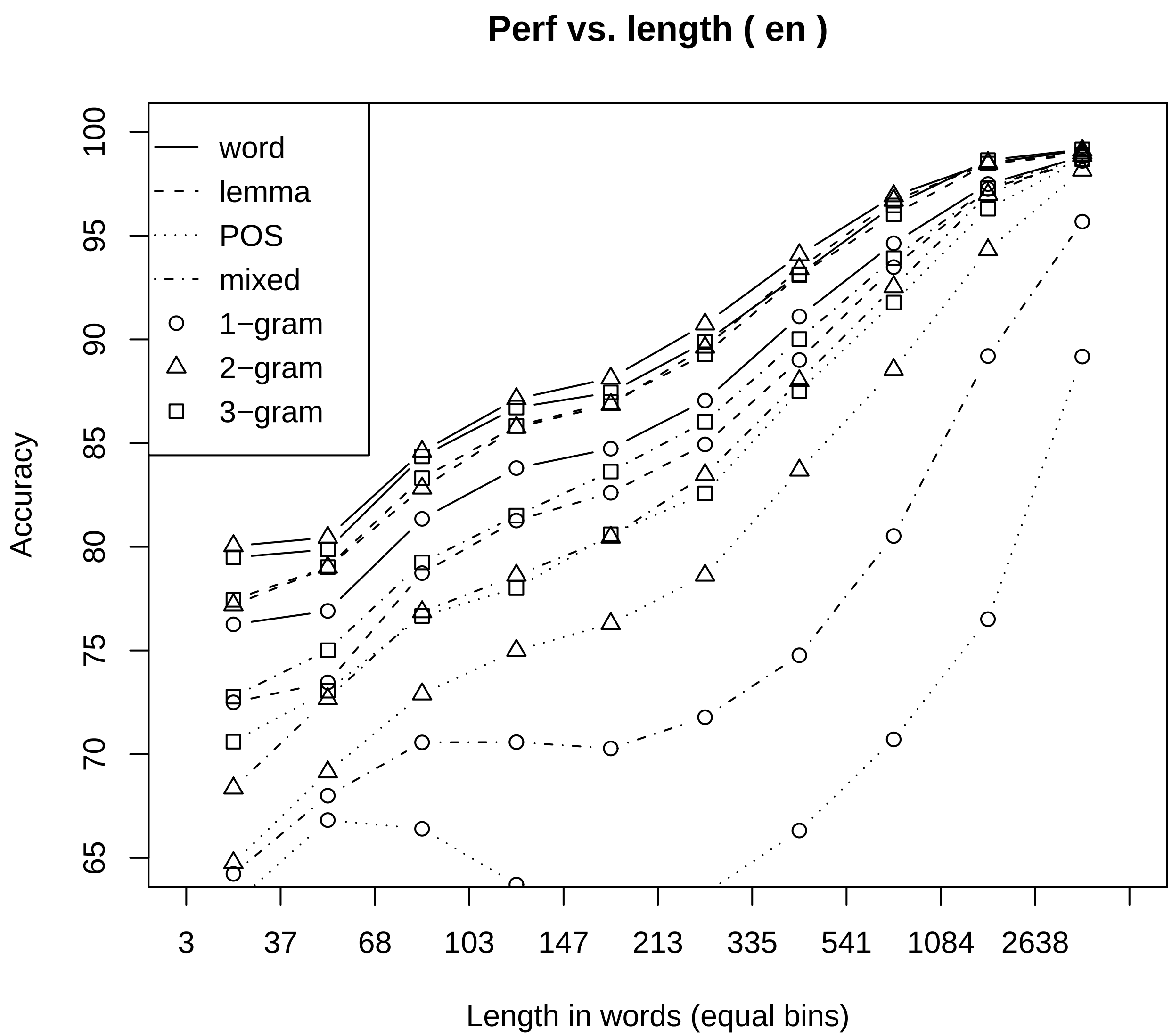
Similar perf. on French, +1-2% for bilingual, same general shape.

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



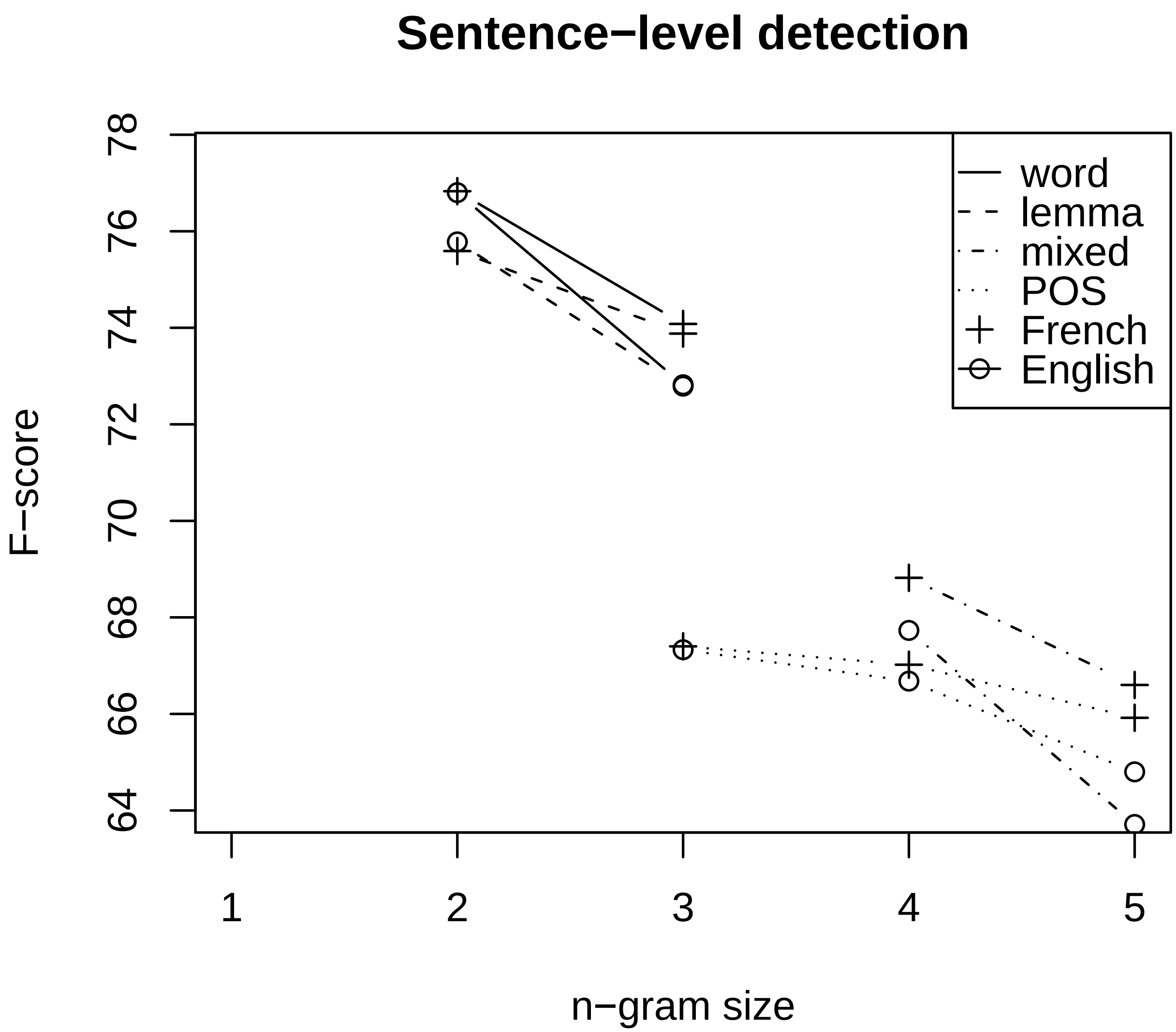
Large range in block length (3-73887 words!).

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”!

eo	fo
couple_of	of_the
alliance_)	mr_.
a_couple	,_the
do_that	in_the
,_canadian	to_the
the_record	,_i
forward_to	._the
,_cpc	)_:
cpc_)	speaker_,
of_us	._i
this_country	:_mr
this_particular	,_and
many_of	._speaker
canadian_alliance	bq_)
across_the	,_bq
out_there	hon_.
the_things	that_the
for_that	on_the

Index

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

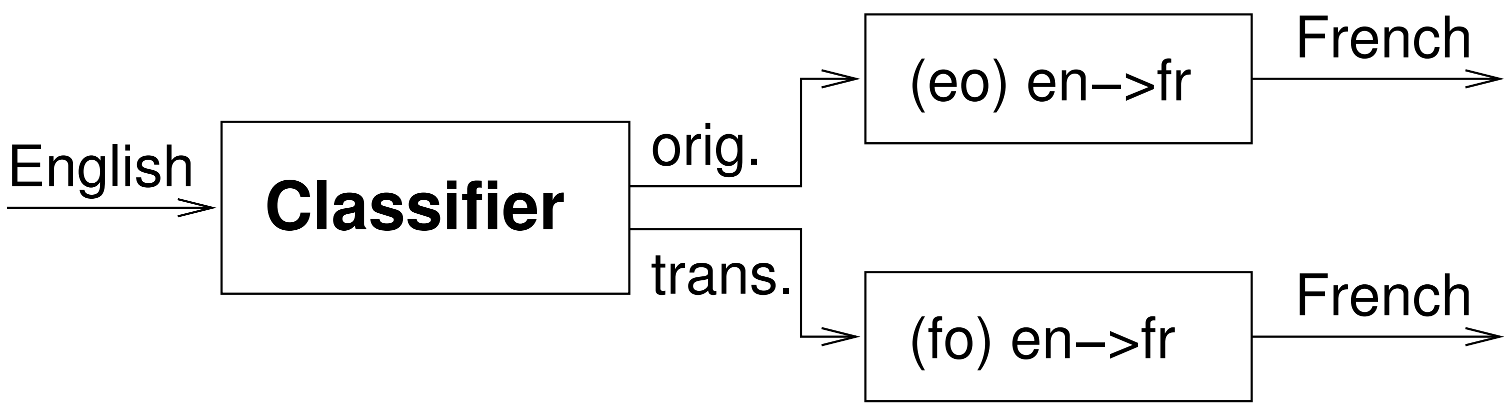
We analyze the impact of translation direction on MT by investigating:

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).

	mx test set		fo test set		eo test set	
Train	fr▷en	en▷fr	fr▷en	en▷fr	fr▷en	en▷fr
mx	36.2	37.1	36.1	37.3	36.1	36.9
fo	31.2	30.8	36.2	36.5	30.5	30.1
eo	36.6	37.8	33.7	36.0	36.8	38.0

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")

Impact of Automatic Detection

Top part is more or less identical to previous table.

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

`SVM`: using SVM prediction, gain is similar.

	Full test set	
	fr→en	en→fr
mx	36.86	37.78
fo	32.00	31.85
eo	37.20	38.23
`SVM`	37.44	38.35
`ref`	37.46	38.35

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

$\Rightarrow$ Detecting original vs. translation provides a small-ish but consistent improvement in translation performance.

$\Rightarrow$ not worth looking for better classifier (for *that* task).

Other uses of translation direction detection?

Index

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