Error Analysis of Error Corrections in Machine Translation

Michel Simard
George Foster, Pierre Isabelle, Roland Kuhn

National Research Council Canada
firstname.lastname@nrc.gc.ca

Workshop Errare 2013, Ermenonville, France

Machine Translation

1949

"Having had considerable exposure to computer design problems during the [Second World] war, and being aware of the speed, capacity, and logical flexibility possible in modern electronic computers, it was very natural for [Mr. Weaver] to think, several years ago, of the possibility that such computers be used for translation."

(Warren Weaver (1949). Translation.)

Machine Translation

1966

*"... when, after 8 years of work, the Georgetown University MT project tried to produce useful output in 1962, they had to resort to **post-editing**. The postedited translation took slightly **longer to do** and was **more expensive** than conventional human translation."*

(ALPAC (1966). Languages and machines: computers in translation and linguistics)

***Postediting** (or post-editing) "is the process of improving a machine-generated translation with a minimum of manual labour".*

(Wikipedia)

Post-editing

2013

- In recent years, MT has improved substantially in terms of quality, cost and availability.
- As a result, many LSPs have now started using MT as a support tool for human translation (post-editing).
- Some users are reporting impressive gains, at least for some application domains and language pairs.

Post-editing Data

- Post-editing has the potential of changing Machine Translation: post-edited translations can be seen as **annotated MT output**, which can be actively used.

> ~~the organization is~~ The authority has actually ~~sentenced~~ been ordered to ~~perform~~ carry out work that ~~he~~ has not been done since 1926 .

- Problem: Inter-annotator agreement [Wisniewski et al., 2013]

> ~~the~~ The organization is actually sentenced to perform work that ~~he~~ it has not done since 1926 .

Post-editing Data

- Post-editing has the potential of changing Machine Translation: post-edited translations can be seen as **annotated MT output**, which can be actively used.

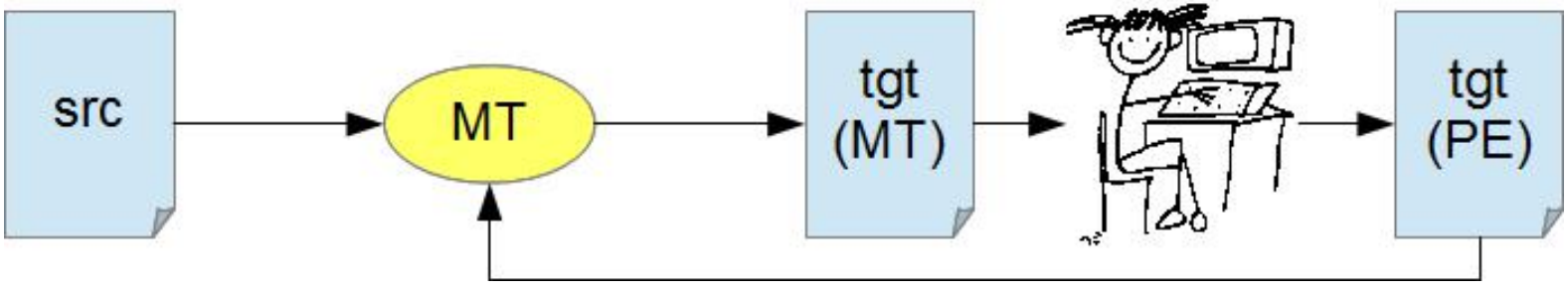
> ~~the organization is~~ The authority has actually ~~sentenced~~ been ordered to ~~perform~~ carry out work that ~~he~~ has not been done since 1926 .

- Problem: Inter-annotator agreement [Wisniewski et al., 2013]

> ~~the~~ The organization is actually sentenced to perform work that ~~he~~ it has not done since 1926 .

Post-editor Feedback

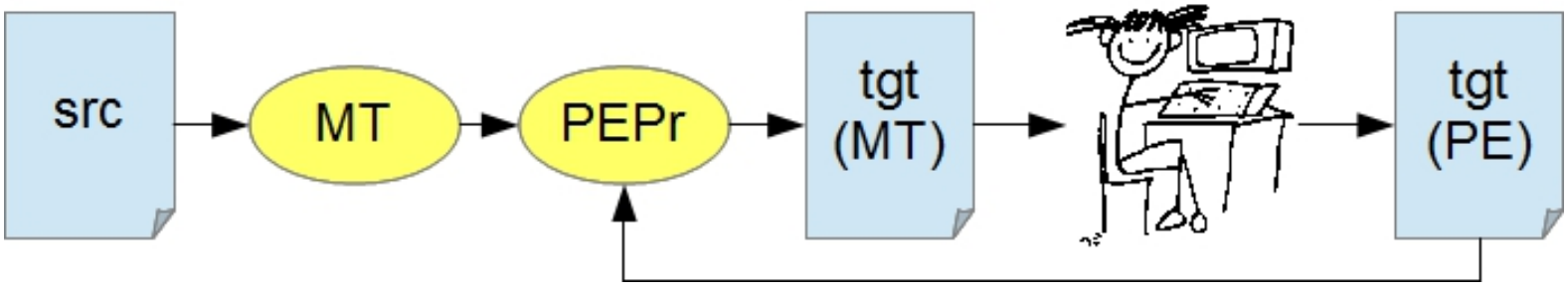
- Application for post-editing data: **improve MT**
- The real challenge: do this in **real-time**, as the post-editor is working.



- One approach: **learn the new translations**, i.e. feed post-edited translations back into the (Statistical) MT system [Nepveu et al., 2004, Levenberg et al., 2010, Hardt and Elming, 2010, Bertoldi et al., 2013]

Post-editor Feedback

- An alternative approach: **learn the corrections**, then perform **Post-edit Propagation** (*PEPr*):

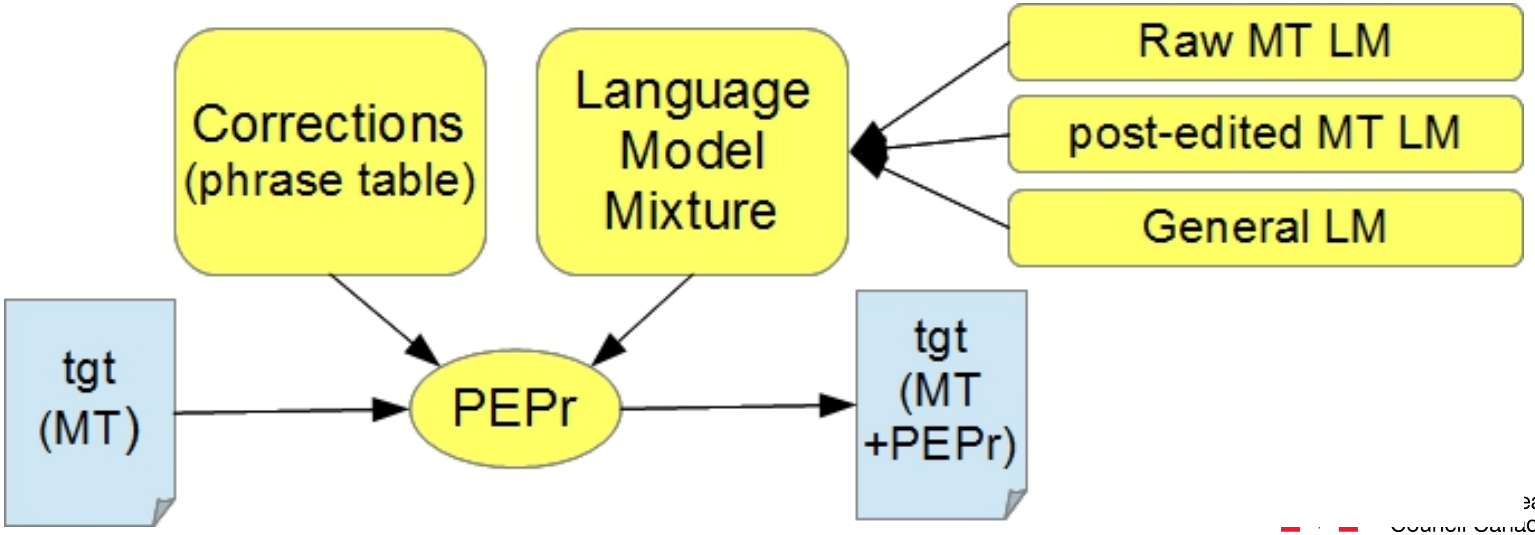


[Simard and Foster, 2013]

- Automatically analyse post-edits as they are produced
- Selectively re-apply learned corrections to further MT output

Post-edit Propagation

- The PEPr system is essentially a **phrase-based SMT** system, with incremental updates
- Learned corrections are stored into a **phrase-table**
- Corrections are performed through **decoding**
- Whether or not a correction is applied depends on its relative frequency and how it scores with the target-language model mixture.



Post-edit Propagation

Post-edited Data

~~the organization is~~ The authority has actually ~~sentenced~~ been ordered to ~~perform~~ carry out work that ~~he~~ has not been done since 1926 .

Learned Corrections

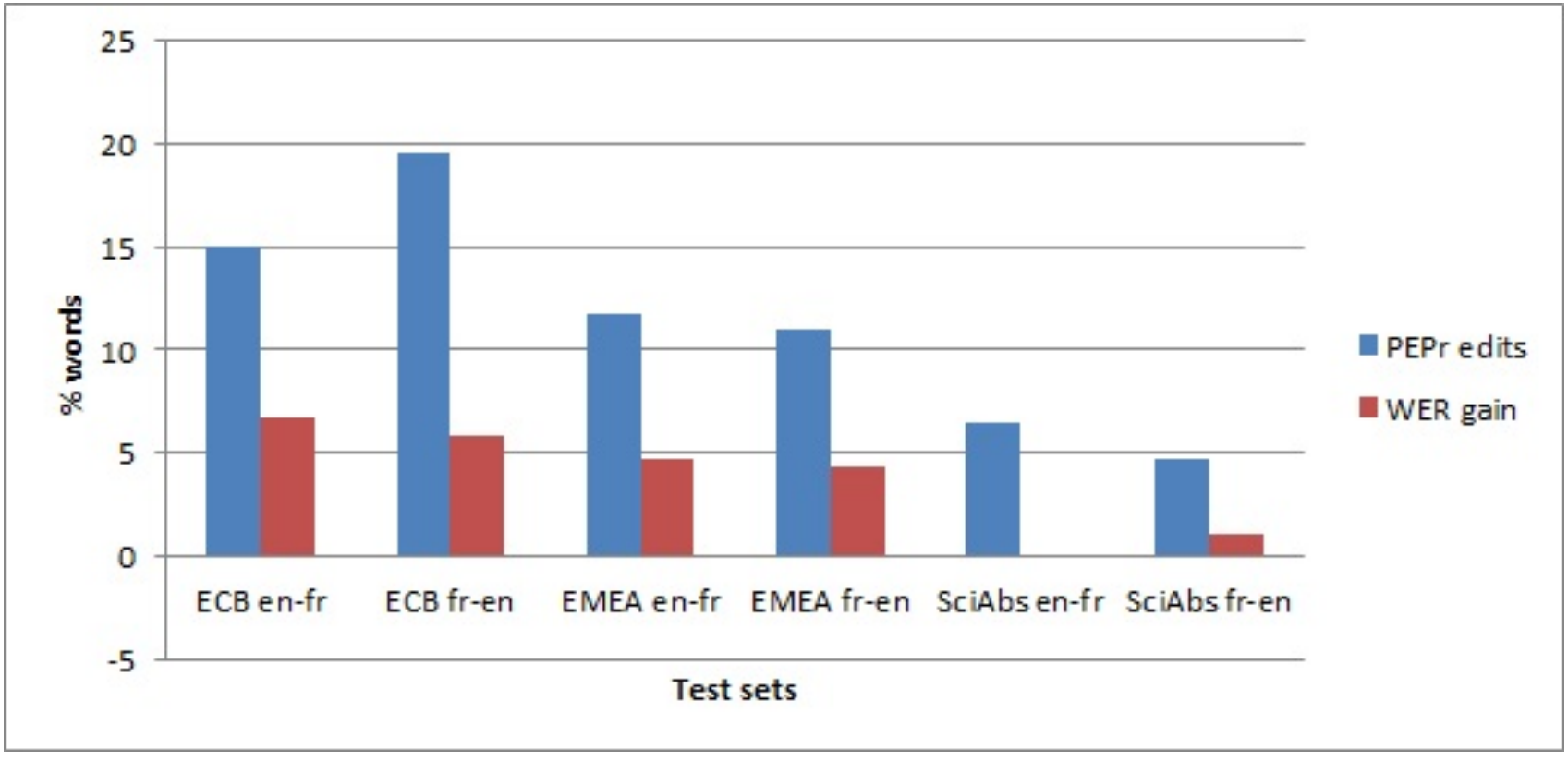
the	→	The
the organization	→	The authority
...		
is	→	has
is actually	→	has actually
...		
to perform work	→	to carry out work
...		
work that	→	work that
...		

Experimental Results

- Test Sets: collections of documents from existing bilingual corpora (*not real post-editing*): ECB, EMEA, Canadian Science Abstracts.
- Raw MT: General-purpose SMT system
- PEPr is effective for technical documents with at least some **internal repetition**.

		WER		PEPr edits
Corpus	Lang.	raw MT	MT + PEPr	(% words)
ECB	en→fr	67.76	61.06	15.06
	fr→en	67.35	61.51	19.54
EMEA	en→fr	67.25	62.60	11.77
	fr→en	59.95	55.57	10.97
Science	en→fr	63.00	63.08	6.42
Abstracts	fr→en	60.36	59.35	4.75

Error Analysis: PEPr gains vs. edits



Error Analysis

- Which corrections are more likely to be **beneficial**?
- Which corrections are **hurting** the most, and **why**?

emea_fr2en_port/test/doc-en_humandocs_PDFs_EPAR_Neorecormon_H-116-PI-en.xml.gz/seg-096					
Source:	Pendant toute la durée du traitement , l' hématocrite ne doit pas dépasser 48 % .				
MT	***PEPr***	***Reference***	***Lev***	***LCS***	***Len***
Throughout the duration of the	Throughout the duration of the	During the entire	+0	+0	+0
treatment ,	therapy ,	treatment	-1	-1	+0
hematocrit	occasions	period	+0	+0	+0
shall	should	, a PCV of 48 % should	+1	+1	+0
not exceed 48 % .	not exceed 48 % .	not be exceeded .	+0	+0	+0
		Total:	**+0**	**+0**	**+0**
		Global:	**+0**	**-1**	**+0**

Error Analysis

- Many PEPr errors are due to **bad alignments**, i.e. cases where our analysis of post-edits breaks down.
  $\rightarrow$ Better alignment methods might help
  $\rightarrow$ Maybe a better idea to **discard segments** that are too heavily post-edited.
- Some corrections don't generalize well:
  - agreement errors
  - function words
  - POS change
  - inserting/removing commas
  - etc.

  Errors that are **contextual** by nature.

Error Analysis

- Many PEPr errors are due to **bad alignments**, i.e. cases where our analysis of post-edits breaks down.
  $\rightarrow$ Better alignment methods might help
  $\rightarrow$ Maybe a better idea to **discard segments** that are too heavily post-edited.
- Some corrections don't generalize well:
  - agreement errors
  - function words
  - POS change
  - inserting/removing commas
  - etc.

  Errors that are **contextual** by nature.

Future Work

- Devise rules-of-thumb for discarding corrections that are potentially bad
- More promising: **Confidence estimation** on individual corrections.
  - Corrections with low confidence can be excluded systematically
  - Alternatively, confidence scores can be used as a selection feature by the PEPr decoder

Thank you!

References I

Bertoldi, N., Cettolo, M., Federico, M., and Kessler, F.-F. B. (2013).

Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation.
In Proc. of MT Summit, Nice, France.

Hardt, D. and Elming, J. (2010).

Incremental Re-training for Post-Editing SMT.
In AMTA.

Levenberg, A., Callison-Burch, C., and Osborne, M. (2010).

Stream-based Translation Models for Statistical Machine Translation.
In NAACL.

Nepveu, L., Lapalme, G., Langlais, P., and Foster, G. (2004).

Adaptive Language and Translation Models for Interactive Machine Translation.
In EMNLP.

Simard, M. and Foster, G. (2013).

PEPr: Post-Edit Propagation Using Phrase-based Statistical Machine Translation.
In Proc. of MT Summit, Nice, France.

Wisniewski, G., Singh, A. K., Segal, N., Neuilly, F., and Yvon, F. (2013).

Design and analysis of a large corpus of post-edited translations: Quality estimation, failure analysis and the variability of post-edition.
In Proc. of MT Summit, Nice, France.