The Effect of Translationese on Statistical Machine Translation

Gennadi Lembersky,
University of Haifa
NICE Systems

EAMT, 2014
Best Thesis Award

Background

Research in Translation Studies indicates that translated texts are different from original ones. Translated texts generally exhibit:

- ***Simplification*** *of the message, the grammar or both* (Al-Shabab, 1996, Laviosa, 1998) ;
- ***Explicitation***, the tendency to spell out implicit utterances that occur in the source text (Blum-Kulka, 1986).
- ***Interference,*** the fingerprints carried over from source- to target-texts (Gellerstam, 1986)

Background

Translated texts can be distinguished automatically from original texts with high accuracy (87% and more)

- *Italian (Baroni & Bernardini, 2006)*
- *Spanish (Ilisei et al., 2010);*
- *English (Koppel & Ordan, 2011; Volansky et al., 2014)*
- *Variety of European Languages (Van Halteren, 1998)*
- *Hebrew (Avner et al, forthcoming)*

Research Question

Human translators produce texts that are different from texts originally written in the target language.

How can we apply these insights from Translation Studies to improve the quality of Statistical Machine Translation?

We investigate the effect of translationese on language and translation models.

Language Models for SMT

Original vs. Translated Texts

Gennadi Lembersky, Noam Ordan and Shuly Wintner. Language Models for Machine Translation: Original vs. Translated Texts. Computational Linguistics38(4):799-825, December 2012.

Research Question

Will using translated texts to compile language models improve the quality of SMT systems?

We investigate this question in three steps:

1. Test the fitness of language models compiled from translated texts to human translations vs. the fitness of LMs compiled from texts written originally in the target language.
2. Test the fitness of language models compiled from texts translated from other languages.
3. Test if language models compiled from translated texts are better for MT than LMs compiled from original texts.

Perplexity Results

French-English	
Original Language	**Perplexity**
English translated from French	**68.37**
Mixture of English translated texts	72.68
English translated from Italian	76.36
English translated from German	81.41
English translated from Dutch	83.55
Original English	**88.31**

- Europarl corpus used to train **6 ENGLISH** equally –sized LMs
- Disjoint portions of Europarl used as the reference set

Perplexity Results

- These results are robust and consistent over:
  - 6 additional language pairs:
    - IT-EN, DE-EN, NL-EN
    - EN-FR, EN-DE
    - HE-EN
  - Different values of N (=1..4) for N-Gram LMs
  - Abstraction Experiments (from eliminating named entities to representing texts solely as POS tag sequences)

MT Results

French - English	
Original Language	**BLEU**
English translated from French	**29.14**
Mixture of English translated texts	28.67
English translated from Italian	28.75
English translated from German	28.01
English translated from Dutch	28.11
Original English	**27.98**

- Human evaluators (MechTurk) also judged that English sentences generated by an MT system whose language model is compiled from translated texts are more fluent than ones generated by a system built with an O-based language model.

Does size matter?

- How much more *original* text do we need to match the performance of a LM trained on *translated* texts?
- To test this we use the Canadian Hansard (French-English bilingual corpus; 80% original English)
  - 3 LMs trained on texts translated from French: 1M words, 5M and 10M
  - 6 LMs trained on English original texts: 1M, 5M, 10M, 25M 50M 100M

Does size matter?

In-domain		
Size	**Original French**	**Original English**
1 M	**33.03**	**31.91**
5 M	**34.25**	**33.27**
10 M	**34.67**	**33.43**
25 M		**33.49**
50 M		**34.29**
100 M		**34.44**

- To achieve the same translation quality, an original English LM must be 10 times larger than a translated LM.

Summary

Practical Outcome:

- Use LMs trained on texts translated from the source language (1 such sentence is worth roughly 10 original sentences)
- Using a mixture of translated texts is the second-best option
- Texts translated from languages that are closely related to the source language are better than other translated texts.

Translation Models

Utilizing the Direction of the Translation

Gennadi Lembersky, Noam Ordan and Shuly Wintner. Improving Statistical Machine Translation by Adapting Translation Models to Translationese. Computational Linguistics 39(4):999-1023, December 2013.

Research Question

Kurokawa et al (2009) show that when translating French into English it is better to use a French-translated-to-English parallel corpus and vice versa.

In case we have parallel corpora translated in both directions, how to build a translation model adapted to the unique properties of the translated text?

Research Question

We investigate this question in three steps:

1. We replicate the results of Kurokawa et al. (2009). We train phrase tables on parallel corpora translated in different directions and apply them to different translation tasks
2. We explain these results by showing that phrase tables built from corpora translated in the 'right' direction are better in terms of various statistical measures.
3. We explore ways to build a translation model adapted to the unique properties of translationese.

Baseline Results

TASK: French-to-English		
Corpus	S->T	T-> S
250K	34.35	31.33
500K	35.21	32.28
750K	36.12	32.90
1M	35.73	33.07
1.5M	36.43	33.73

TASK: English-to-French		
Corpus	S->T	T->S
250K	27.74	26.58
500K	29.15	27.19
750K	29.43	27.63
1M	29.94	27.88
1.5M	29.89	27.83

- S->T (source-to-target) – 'right' direction
- T->S (target-to-source) – 'wrong' direction

Analysis of the Phrase Tables

TASK: French-to-English				
Corpus	S->T		T-> S	
	CovLen	CovEnt	CovLen	CovEnt
250K	2.44	0.36	2.25	0.45
500K	2.64	0.35	2.42	0.43
750K	2.77	0.35	2.53	0.43
1M	2.85	0.34	2.61	0.42
1.5M	2.97	0.33	2.71	0.41

- Given text in the source language:
  - **CovLen** finds the average phrase length of the minimal covering set of the text using source phrases from a particular phrase-table.
  - **CovEnt** searches for the covering set of the text that minimizes the average entropy of the source phrases in the covering set.

Translation Model Adaptation

- **Goal**: Given any bi-text comprising S->T and T->S subsets, improve translation quality by taking advantage of information pertaining to the direction of translation.
- **Techniques**:
  - **Union** – simple concatenation between corpora
  - **Two phrase-tables** – train a phrase table for each subset and give MOSES two phrase tables
  - **Phrase table interpolation** – using perplexity minimization following Sennrich (2012).
  - Add **new feature** in the phrase table that pertains to the direction of translation

"Additional Feature" Results

TASK: French-to-English			
System	S->T to T->S ratio		
	1:1	1:2	2:1
UNION	35.27	35.36	35.94
CrEnt	35.54	35.45	**36.75**
PplRatio	**35.59**	**35.78**	36.22

TASK: English-to-French			
System	S->T to T->S ratio		
	1:1	1:2	2:1
UNION	29.27	29.44	30.01
CrEnt	29.47	29.45	**30.44**
PplRatio	**29.65**	**29.62**	30.34

- **CrEnt** - the cross-entropy of each target phrase with respect to a language model of translated texts.
- **PplRatio** - the ratio between the perplexity of a target phrase with respect to an "original" language model and its perplexity with respect to a "translated" one Based on Moore and Lewis (2010).

Summary

When machine translation meets Translation Studies

1. MT results improve.
2. Pending hypotheses in translation studies are tested experimentally in a more rigorous way.
3. SMT becomes more human.

Future cooperation between these two disciplines is likely to yield beneficial insights for both of them.

Thank You!