THE FEATURES OF TRANSLATIONESE
BETWEEN HUMAN AND MACHINE TRANSLATION

Shuly Wintner

Department of Computer Science
University of Haifa
Haifa, Israel
shuly@cs.haifa.ac.il

ISCOL 2014

University of Haifa, 7 September 2014
Introduction

Original or Translation?

Example (O or T?)

The Features of Translationese

Example (T or O?)

© Shuly Wintner (University of Haifa)

ISCOL 2014
**TRANSLATIONESE**

**THE LANGUAGE OF TRANSLATED TEXTS**

- Translated texts differ from original ones
- The differences do not indicate poor translation but rather a statistical phenomenon, **translationese** (*Gellerstam, 1986*)
- **Toury (1980, 1995)** defines two **laws of translation**:  
  - **THE LAW OF INTERFERENCE** Fingerprint of the source text that are left in the translation product  
  - **THE LAW OF GROWING STANDARDIZATION** Effort to standardize the translation product according to existing norms in the target language and culture
TRANSLATIONUNIVERSALS (Baker, 1993)

“features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems”

SIMPLIFICATION (Blum-Kulka and Levenston, 1978, 1983)

EXPLICITATION (Blum-Kulka, 1986)

NORMALIZATION (Chesterman, 2004)
COMPUTATIONAL INVESTIGATION OF TRANSLATIONSE

- Translated texts exhibit lower lexical variety (type-to-token ratio) than originals (Al-Shabab, 1996)
- Their mean sentence length and lexical density (ratio of content to non-content words) are lower (Laviosa, 1998)
- Corpus-based evidence for the simplification hypothesis (Laviosa, 2002)
Methodology

- Corpus-based approach
- Text classification with machine-learning techniques
- Feature design
- Evaluation
IDENTIFYING TRANSLATIONSE
USING TEXT CLASSIFICATION

- Baroni and Bernardini (2006)
- van Halteren (2008)
- Kurokawa et al. (2009)
- Ilisei et al. (2010); Ilisei and Inkpen (2011); Ilisei (2013)
- Koppel and Ordan (2011)
- Popescu (2011)
**Research Contributions**

- Understanding the features of translationese; testing Translation Studies hypotheses *(Volansky et al., Forthcoming; Avner et al., Forthcoming)*
- Robust classification of translationese *(Twitto-Shmuel et al., Forthcoming)*
- Language models for statistical machine translation *(Lembersky et al., 2011, 2012b)*
- Translation models for statistical machine translation *(Kurokawa et al., 2009; Lembersky et al., 2012a, 2013)*
- Automatic detection of **machine** translated texts *(Aharoni et al., 2014)*
- Identifying the first language of non-native writers *(Tsvetkov et al., 2013)*

Goal: test Translation Studies hypotheses using classification as a methodology

Experimental setup: EUROPARL, 4M tokens in English (O) and 400K tokens translated from each of ten European languages (T)

After tokenization, the corpus is partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary)

Classification with Weka *(Hall et al., 2009)*, using SVM with a default linear kernel
Hypotheses

Simplification  Rendering complex linguistic features in the source text into simpler features in the target (Blum-Kulka and Levenston, 1983; Vanderauwerea, 1985; Baker, 1993)

Explanation  The tendency to spell out in the target text utterances that are more implicit in the source (Blum-Kulka, 1986; Øverås, 1998; Baker, 1993)

Normalization  Efforts to standardize texts (Toury, 1995), “a strong preference for conventional grammaticality” (Baker, 1993)

Interference  The fingerprints of the source language on the translation output (Toury, 1979)
Features Should...

1. Reflect frequent linguistic characteristics we would expect to be present in the two types of text.
2. Be content-independent, indicating formal and stylistic differences between the texts that are not derived from differences in contents, domain, genre, etc.
3. Be easy to interpret, yielding insights regarding the differences between original and translated texts.
FEATURES

SIMPLIFICATION  Type-token ratio, Mean word length, Syllable ratio, Mean sentence length, Lexical density, Mean word rank, Most frequent words

EXPLICITATION  Explicit naming, Single naming, Mean multiple naming, Cohesive markers

NORMALIZATION  Repetitions, Contractions, Average PMI, Threshold PMI

INTERFERENCE  POS n-grams, Character n-grams, Prefixes and suffixes, Contextual function words, Positional token frequency

MISCELLANEOUS  Function words, Pronouns, Punctuation, Ratio of passive forms, Token unigrams and bigrams
### Results: Sanity Check

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanity</td>
<td>Token unigrams</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Token bigrams</td>
<td>100</td>
</tr>
</tbody>
</table>
## Results: Simplification

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplification</td>
<td>TTR (1)</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>TTR (2)</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>TTR (3)</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Mean word rank (1)</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Mean word rank (2)</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>$N$ most frequent words</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Mean word length</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Syllable ratio</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>Lexical density</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Mean sentence length</td>
<td>65</td>
</tr>
</tbody>
</table>
### Results: Explicitation

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicitation</td>
<td>Cohesive Markers</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Explicit naming</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Single naming</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Mean multiple naming</td>
<td>54</td>
</tr>
</tbody>
</table>
## RESULTS: NORMALIZATION

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalization</td>
<td>Repetitions</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>Contractions</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Average PMI</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Threshold PMI</td>
<td>66</td>
</tr>
</tbody>
</table>
## Results: Interference

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interference</td>
<td>POS unigrams</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>POS bigrams</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>POS trigrams</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Character unigrams</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Character bigrams</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Character trigrams</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Prefixes and suffixes</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Contextual function words</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Positional token frequency</td>
<td>97</td>
</tr>
</tbody>
</table>
## Results: Reduced Parameter Space

(300 Most Frequent Features)

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interference</td>
<td>POS bigrams</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>POS trigrams</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>Character bigrams</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Character trigrams</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>Positional token frequency</td>
<td>93</td>
</tr>
</tbody>
</table>
### Results: Miscellaneous

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miscellaneous</td>
<td>Function words</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>Punctuation (1)</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Punctuation (2)</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Punctuation (3)</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Pronouns</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Ratio of passive forms to all verbs</td>
<td>65</td>
</tr>
</tbody>
</table>
CONCLUSION

- Machines can accurately identify translated texts
- The best performing features are those that attest to the ‘fingerprints’ of the source on the target
- Interference by its nature is a pair-specific phenomenon
- Translation “universals” should be reconsidered. Not only are they dependent on genre and register, they also vary greatly across different pairs of languages
- Ideally, such claims should be studied using a comparable corpus
Other Contributions

Other Contributions


FUTURE DIRECTIONS

- Identification of translationese at the sentence-pair level
- The features of **machine** translationese
- More applications to machine translation
ACKNOWLEDGEMENTS

- Noam Ordan, Vered Volansky, Ehud Alexander Avner, Naama Twitto, Gennadi Lembersky, Moshe Koppel
- Israel Ministry of Science and Technology
BIBLIOGRAPHY


Gideon Toury. In Search of a Theory of Translation. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv, 1980.


