Translationese: Between Human and Machine Translation

Shuly Wintner
Department of Computer Science, University of Haifa
Mount Carmel, Haifa, Israel
shuly@cs.haifa.ac.il
http://www.cs.haifa.ac.il/~shuly

Brief Description

Translated texts, in any language, have unique characteristics that set them apart from texts originally written in the same language. Translation Studies is a research field that focuses on investigating these characteristics. Until recently, research in machine translation (MT) has been entirely divorced from translation studies. The main goal of this tutorial is to introduce some of the findings of translation studies to researchers interested mainly in machine translation, and to demonstrate that awareness to these findings can result in better, more accurate MT systems.

First, we will survey some theoretical hypotheses of translation studies. Focusing on the unique properties of *translationese* (the sub-language of translated texts), we will distinguish between properties resulting from *interference* from the source language (the so-called "fingerprints" of the source language on the translation product) and properties that are source-language-independent, and that are presumably universal. The latter include phenomena resulting from three main processes: *simplification*, *standardization* and *explicitation*. All these phenomena will be defined, explained and exemplified.

Then, we will describe several works that use standard (supervised and unsupervised) text classification techniques to distinguish between translations and originals, in several languages. We will focus on the features that best separate between the two classes, and how these features corroborate some (but not all) of the hypotheses set forth by translation studies scholars.

Next, we will discuss several computational works that show that awareness to translationese can improve machine translation. Specifically, we will show that language models compiled from translated texts are more fitting to the reference sets than language models compiled from originals. We will also show that translation models compiled from texts that were (manually) translated from the source to the target are much better than translation models compiled from texts that were translated in the reverse direction. We will briefly discuss how translation models can be adapted to better reflect the properties of translationese.

Finally, we will touch upon some related issues and current research directions. For example, we will discuss recent work that addresses the identification of the source language from which target language texts were translated. We will show that native language identification (in particular, of language learners) is a closely related task to the identification of translationese. Time permitting, we will also discuss work aimed at distinguishing between native and (advanced, fluent) non-native speakers.

Outline

- Translation Studies hypotheses
  - Simplification
  - Explicitation
  - Normalization
  - Interference
- Identification of translationese
  - Text classification

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tutorial Abstracts,
pages 18–19, Osaka, Japan, December 11-17 2016.

  - Features
  - Supervised classification
  - Unsupervised classification

- Relevance for machine translation
  - Improving language models
  - Improving translation models

- Related issues
  - Identification of the source language of translations
  - Native language identification
  - Distinguishing between native and non-native speakers

Instructor

Shuly Wintner is a professor of computer science at the University of Haifa, Israel. His research spans various areas of computational linguistics and natural language processing, including formal grammars, morphology, syntax, language resources, and translation. He served as the editor-in-chief of Springer's Research on Language and Computation, a program co-chair of EACL-2006, and the general chair of EACL-2014. He was among the founders, and twice (6 years) the chair, of ACL SIG Semitic. Currently, he serves as the Head of the Department of Computer Science in Haifa. Shuly has an extensive teaching experience, including tutorials at EACL-2012, ICGI-2012, NAACL-2004, MT-Summit 2003 and COLING-2000; five ESSLLI courses; three courses at the International PhD School in Formal Languages and Applications; and two at the Erasmus Mundus Master course in Language and Communication Technology.