Mixed up with Machine Translation: Multi-word Units Disambiguation Challenge

Anabela BARREIRO[1], Annibale ELIA[2], Johanna MONTI[2] and Mario MONTELEONE[2]
barreiro_anabela@hotmail.com, elianni@tin.it, jmont@tin.it, mmonteleone@unisa.it,

[1] Centro de Linguística da Universidade do Porto, Via Panorâmica, s/n, 4150 – 564 Porto, Portugal
[2] Dip. di Scienze della Comunicazione, Università degli Studi di Salerno, Via Ponte Don Melillo 84084, Fisciano, Italia

[forthcoming in: Proceedings of the Translating and the Computer ASLIB Conference 2010]

Abstract

With the rapid evolution of the Internet, translation has become part of the daily life of ordinary users, not only of professional translators. Machine translation has evolved along with different types of computer-assisted translation tools. Qualitative progress has been made in the field of machine translation, but not all problems have been solved. The current times are auspicious for the development of more sophisticated evaluation tools that measure the performance of specific linguistic phenomena. One problem in particular, namely the poor analysis and translation of multi-word units, is an arena where investment in linguistic knowledge systems with the goal of improving machine translation would be beneficial. This paper addresses the difficulties multi-word units present to machine translation, by comparing translations performed by systems adopting different approaches to machine translation. It proposes a solution for improving the quality of the translation of multi-word units by adopting a methodology that combines Lexicon Grammar resources with OpenLogos lexical resources and semantico-syntactic rules. Finally, it discusses how an ideal machine translation evaluation tool might look to correctly evaluate the performance of machine translation engines with regards to multi-word units and thus to contribute to the improvement of translation quality.

Introduction

The Internet has helped machine translation to become increasingly popular within the general public. Today millions of Internet users take advantage of machine translation to quickly obtain information on the contents of a text or a web page written in a foreign language, to exchange information in real-time, to retrieve information in unknown languages, or even to produce publishable translations. Most recently, machine translation is used for dissemination purposes in online collaborative translation environments (Monti, forthcoming). This unpredictably quick turn on machine translation usability complements traditional uses as the ones described in Hutchins (2005), where the challenge of producing high quality translations was big, but more controllable. The world of machine translation has changed forever: the spectrum of language to be translated by machines is now broadening and more complex, less controlled and more idiomatic. Considerable progress has also been made in qualitative terms because of the availability and use of large parallel corpora, the development of knowledge bases, the adoption of statistical models, and the integration with various computer

[1] Anabela Barreiro is author of abstract, introduction and section 4, Annibale Elia is author of section 1, Mario Monteleone is author of section 2, Johanna Monti is author of sections 3 and 5 and conclusions.

assisted translation tools, particularly with translation memories. However, despite recent significant progress, lexical problems still represent a critical area in machine translation, and among lexical problems, multi-word units, are particularly difficult to be processed by machine translation systems.

The aim of this paper is to provide evidence of the shortcomings of existing machine translation systems with reference to the processing of multi-word units, and in the line of thought of evaluation proposed by (Barreiro, 2008) suggest a systematic qualitative evaluation of different linguistic phenomena, starting with multi-word units with different degrees of variability. The paper points out benefits, strengths and weaknesses of distinct machine translation approaches and discusses the usage of combined Lexicon-Grammar lexical resources and OpenLogos lexical resources together with semantico-syntactic rules (SEMTAB rules) as a possible solution to overcome machine translation limitations with regard to the automated processing and translation of multi-word units. We propose that, for a fair machine translation evaluation activity, there is the need for a serious joint qualitative evaluation of the systems to balance with the numerous quantitative evaluations that have taken place in the latest years by automated evaluation tasks, including BLEU, NIST and METEOR, which we consider insufficient to measure translation accuracy and linguistic quality. We propose that qualitative evaluation will be made with the aid of a new machine translation evaluation tool, conceptually inspired in existing tools, such as METRA and TrAva (Santos et al., 2004) (Sarmento, 2007) created by the Linguateca project. This paper presents the results of a research based on the translation of sentences containing multi-word units from English into Italian of a non-specialised text corpus. Section 1 presents the notion of multi-word unit in the framework of the Lexicon Grammar theory. Section 2 analyses how multi-word units are processed using the state-of-the-art machine translation technology. Section 3 discusses several examples of lexical ambiguities concerning multi-word units in the translations performed by a statistical machine translation system and a rule-based machine translation system, and highlights, analyzes and discusses how two machine translation systems of a different conceptual nature perform with regards to different types of multi-word unit. Section 4 discusses the possibility of using syntactic-semantic rules in order to obtain better translation quality results. Section 5 discusses how the "ideal" machine translation evaluation tool would look like to correctly evaluate the performance of machine translation engines with regards to multi-word units. Section 6 presents the conclusions.