Accepted Manuscript

Machine translation using semantic web technologies: A Survey

Diego Moussallem, Matthias Wauer, Axel-Cyrille Ngonga Ngomo

PII: S1570-8268(18)30030-1
DOI: https://doi.org/10.1016/j.websem.2018.07.001
Reference: WEBSEM 464

To appear in: *Web Semantics: Science, Services and Agents on the World Wide Web*

Received date : 29 January 2018
Revised date : 4 June 2018
Accepted date : 13 July 2018

Please cite this article as: D. Moussallem, M. Wauer, A.N. Ngomo, Machine translation using semantic web technologies: A Survey, *Web Semantics: Science, Services and Agents on the World Wide Web* (2018), https://doi.org/10.1016/j.websem.2018.07.001

Machine Translation using Semantic Web Technologies: A Survey

Diego Moussallem[a,b,*], Matthias Wauer[a], Axel-Cyrille Ngonga Ngomo[b]

[a] *University of Leipzig*
AKSW Research Group
Department of Computer Science
Augustusplatz 10, 04109 Leipzig, Germany
[b] *University of Paderborn*
Data Science Group
Pohlweg 51, D-33098 Paderborn, Germany

Abstract

A large number of machine translation approaches have recently been developed to facilitate the fluid migration of content across languages. However, the literature suggests that many obstacles must still be dealt with to achieve better automatic translations. One of these obstacles is ambiguity. A promising way of overcoming this problem is using Semantic Web technologies. This article presents the results of a systematic review of machine translation approaches that rely on Semantic Web technologies for translating texts. Overall, our survey suggests that while Semantic Web technologies can enhance the quality of machine translation outputs for various problems, the combination of both is still in its infancy.

Keywords: machine translation, semantic web, ontology, linked data, multilinguality, knowledge graphs

1. Introduction

Alongside increasing globalization comes a greater need for readers to understand texts in languages foreign to them. For example, approximately 48% of the pages on the Web are not available in English[1]. The technological progress of recent decades has made both the distribution and access to content in different languages ever simpler. Translation aims to support users who need to access content in a language in which they are not fluent [1, 2].

However, translation is a difficult task due to the complexity of natural languages and their structure [3]. In addition, manual translation does not scale to the magnitude of the Web. One remedy for this problem is Machine Translation (MT). The main goal of MT is to enable people to assess content in languages other than the languages in which they are fluent [4]. From a formal point of view, this means that the goal of MT is to transfer the semantics of text from an input language to an output language [5]. At the time of writing, large information portals such as Google[2] or Bing[3] already offer MT services that are widely used.

Although MT systems are now popular on the Web, they still generate a large number of incorrect translations. Recently, Popović [6] has classified five types of errors that still remain in MT systems. According to research, the two main faults that are responsible for 40% and 30% of problems respectively, are reordering errors and lexical and syntactic ambiguity. Thus, addressing these barriers is a key challenge for modern translation systems.

A large number of MT approaches have been developed over the years that could potentially serve as a remedy. For instance, translators began by using methodologies based on linguistics which led to the family of Rule-Based Machine Translation (RBMT). However, RBMT systems have a critical drawback in their reliance on manually crafted rules, thus making the development of new translation modules for different languages even more difficult [7, 8].

Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT) were developed to deal with the scalability issue in RBMT [9], a necessary characteristic of MT systems that must deal with data at Web scale. Presently, these approaches have begun to address the drawbacks of rule-based approaches. However, some problems that had already been solved for linguistics based methods reappeared. The majority of these problems are connected to the issue of ambiguity, including syntactic and semantic variations [2].

Subsequently, RBMT and SMT have been combined in order to resolve the drawbacks of these two families of approaches. This combination of methods is called hybrid MT. Although hybrid approaches have been achieving good results, they still suffer from some RBMT problems [10–12], for example, the big effort of adding new rules for handling a given syntax divergence. Nowadays, a novel SMT paradigm has arisen called Neural Machine Translation (NMT) which relies on Neural Network (NN) algorithms. NMT has been achieving im-

*Principal corresponding author
Email addresses: moussallem@informatik.uni-leipzig.de (Diego Moussallem), wauer@informatik.uni-leipzig.de (Matthias Wauer), axel.ngonga@upb.de (Axel-Cyrille Ngonga Ngomo)

[1] https://www.internetworldstats.com/stats7.htm
[2] http://translate.google.com.br/about/
[3] http://www.bing.com/translator/help/