Free/Open-Source Machine Translation for the Low-Resource Languages of Spain

Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, 03690 Sant Vicent del Raspeig, Spain
Prompsit Language Engineering, 03202 Elx, Spain

Abstract

While machine translation has historically been rule-based, that is, based on dictionaries and rules written by experts, most present-day machine translation is corpus-based. In the last few years, statistical machine translation, the dominant corpus-based approach, has been displaced by neural machine translation in most applications, in view of the better results reported, particularly for languages with very different syntax. But both statistical and neural machine translation need to be trained on large amounts of parallel data, that is, sentences in one language carefully paired with their translations in their other language, and this is a resource that may not be available for some low-resource languages. While some of the languages of Spain may be considered to be reasonably endowed with parallel corpora connecting them to Spanish or even to English – Basque, Catalan, Galician –, and are well-served with machine translation systems, there are many other languages which cannot afford them such as Aranese Occitan, Aragonese, or Asturian/Leonese. Fortunately, languages in this last group belong to the Romance language family, as Spanish does, and this makes translation from and into Spanish under a rule-based paradigm the only feasible approach. After describing briefly the main machine translation paradigms, I will describe the Apertium free/open-source rule-based machine translation platform, which has been used to build machine translation systems for these low-resource languages of Spain, indeed, sometimes the only ones available. The free/open-source setting has made linguistic data for these languages available for anyone in their linguistic communities to build other linguistic technologies for these low-resourced languages. For example, the Apertium family of bilingual and monolingual data has been converted into RDF and they have been made accessible on the Web as linked data.

2012 ACM Subject Classification Applied computing → Language translation

Keywords and phrases free/open-source, machine translation, languages of Spain, low-resource machine translation

Digital Object Identifier 10.4230/OASIcs.LDK.2021.3

Category Invited Talk

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch; Article No. 3; pp. 3:1–3:1
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany