Statistical Machine Translation tools in $R$

Neda Daneshgar$^{1,2}$, Majid Sarmad$^{1,3,*}$

1. Ferdowsi University of Mashhad, Faculty of Mathematical Sciences
2. PhD Student in Statistics (Inference)
3. Assistant Professor in Statistics (Statistical Computations)
*Contact author: sarmad@um.ac.ir

Keywords: SMT, MPI, IBM Models

Statistical Machine Translation (SMT) is a method to work on parallel corpuses to translate a text in another language (as much as fluent) without any knowledge of both languages. The book is written by Koehn (University of Edinburgh) explains IBM models (including algorithms) to achieve the goal. There are some tools to work on parallel corpuses but no package in $R$. We have started to write the package in $R$ and IBM model-I is roughly completed and works fine. The most important problem is that parallel corpuses are big (better to say huge), then we need to apply some techniques to handle the large objects with a reasonable speed. MPI (Message-Passing Interface) packages in $R$ would help us to make the codes for a parallel processing. Furthermore, we are going to compare our speed with the other tools like Moses and Giza++.

Reference