# Comparing Deep Architectures for Selecting Optimal Machine Translation

**Authors :** Despoina Mouratidis, Katia Lida Kermanidis

**Abstract :** Machine translation (MT) is a very important task in Natural Language Processing (NLP). MT evaluation is crucial in MT development, as it constitutes the means to assess the success of an MT system, and also helps improve its performance. Several methods have been proposed for the evaluation of (MT) systems. Some of the most popular ones in automatic MT evaluation are score-based, such as the BLEU score, and others are based on lexical similarity or syntactic similarity between the MT outputs and the reference involving higher-level information like part of speech tagging (POS). This paper presents a language-independent machine learning framework for classifying pairwise translations. This framework uses vector representations of two machine-produced translations, one from a statistical machine translation model (SMT) and one from a neural machine translation model (NMT). The vector representations consist of automatically extracted word embeddings and string-like language-independent features. These vector representations used as an input to a multi-layer neural network (NN) that models the similarity between each MT output and the reference, as well as between the two MT outputs. To evaluate the proposed approach, a professional translation and a "ground-truth" annotation are used. The parallel corpora used are English-Greek (EN-GR) and English-Italian (EN-IT), in the educational domain and of informal genres (video lecture subtitles, course forum text, etc.) that are difficult to be reliably translated. They have tested three basic deep learning (DL) architectures to this schema: (i) fully-connected dense, (ii) Convolutional Neural Network (CNN), and (iii) Long Short-Term Memory (LSTM). Experiments show that all tested architectures achieved better results when compared against those of some of the well-known basic approaches, such as Random Forest (RF) and Support Vector Machine (SVM). Better accuracy results are obtained when LSTM layers are used in our schema. In terms of a balance between the results, better accuracy results are obtained when dense layers are used. The reason for this is that the model correctly classifies more sentences of the minority class (SMT). For a more integrated analysis of the accuracy results, a qualitative linguistic analysis is carried out. In this context, problems have been identified about some figures of speech, as the metaphors, or about certain linguistic phenomena, such as per etymology: paronyms. It is quite interesting to find out why all the classifiers led to worse accuracy results in Italian as compared to Greek, taking into account that the linguistic features employed are language independent.

**Keywords :** machine learning, machine translation evaluation, neural network architecture, pairwise classification
**Conference Title :** ICMLA 2020 : International Conference on Machine Learning and Applications
**Conference Location :** Copenhagen, Denmark
**Conference Dates :** June 11-12, 2020