Thursday 7 February 2019
11h00 – 12h00
UT3 Paul Sabatier, IRIT, Salle du Conseil

Radu TUDOR IONESCU
University of Bucharest (Roumanie)

Usage patterns of non-native language speakers discovered by string kernels for native language identification

Abstract: Recently, an approach that uses only character p-grams as features has been proposed for the task of native language identification (NLI). The approach obtained state-of-the-art results by combining several string kernels using multiple kernel learning. A broad set of native language identification experiments are presented to compare the string kernels approach with other state-of-the-art methods. The empirical results obtained in the experiments indicate that the proposed approach achieves state-of-the-art performance in NLI. To gain additional insights about the string kernels approach, the features selected by the classifier as being more discriminating are analyzed in this presentation. The analysis also offers information about localized language transfer effects, since the features used by the proposed model are p-grams of various lengths. The features captured by the model typically include stems, function words, word prefixes and suffixes, which have the potential to generalize over purely word-based features. By analyzing the discriminating features, this presentation offers insights into two kinds of language transfer effects, namely word choice (lexical transfer) and morphological differences.