Abstract

The world has a data revolution, which caused a huge volume of data stored in documents in different languages. This creates an important demand for developing Cross-lingual resources to serve NLP applications, in order to understand, retrieve, translate, summarize such large amounts of texts. Thesaurus and WordNet are examples for cross-lingual resources, which become core components in modern NLP applications, specially to support multilingualism.

Although there are several thesauruses for Arabic, the majority are messy - instead of providing accurate sets of synonyms for a given word, they provide "near" synonyms and general/specific words. For example, according to the Google's Arabic thesaurus, the synonyms of the words " دولة " are {بلد, قطر, وطن, الريف, ريف}. Here the underlined words are wrong, as they are not really synonyms.

In this thesis we build an Arabic thesaurus automatically and map this thesaurus to the English WordNet. That is, the result will be a set of Arabic synsets mapped into WordNet synsets, as $\{a_1, a_2, \ldots, a_n\} := \{wn_1, wn_2, \ldots, wn_m\}$. To do this, we will first implement SynsetGenerator algorithm for generating multilingual thesauruses which requires a set of Arabic-English bilingual dictionaries as input, then we evaluate our results and link the generated results with WordNet using Cosine similarity approach.