

Lecture Summary:

- The lecture is about Statistical Transliteration for English-Arabic Cross Language Information Retrieval. The Out of vocabulary (OOV) words is a common source of errors in cross language information retrieval. We talk about TRANSLITERATION METHOD which was used for the Transliteration. The transliteration models were evaluated in two different ways: The first evaluation uses a measure of translation accuracy which measures the correctness of transliterations generated by the models, using the spellings found in the AFP corpus as the standard for correct spelling. The second kind of evaluation uses a cross language information retrieval task and looks at how retrieval performance changes as a result of including transliterations inquiry translations. Finally, we talk about how the results of there experiments support the following generalizations: Good quality transliteration models can be generate automatically from reasonably small data sets. A hand-crafted model performs slightly better than the automatically-trained model. The quality of the source of training data affects the accuracy of the model. Context dependency is important for the transliteration of English words. The selected n-gram model is more accurate than the unigram model. Results of the IR evaluation confirm that transliteration can improve cross-language IR. However, it is not a good strategy to transliterate names that are already translated in the dictionary.

REFERENCES

- Al Misbar. http://www.almisbar.com/salam_trans.html
- Arabic Proper Names Dictionary from NMSU. <http://crl.nmsu.edu/~ahmed/downloads.html>
- Arbabi, Mansur, Scott M. Fischthal, Vincent C. Cheng, and Elizabeth Bar. 1994. Algorithms for Arabic name transliteration. IBM Journal of research and Development, 38(2):183-193.
- Automatically-trained Transliteration Model. http://www.cs.umass.edu/~nasreen/automatic_model.txt
- GIZA++. <http://www-i6.informatik.rwthachen.de/Colleagues/och/software/GIZA++.html>

Obstacles that I have faced:

Actually I face a problem with some terms like GIZA++.

Questions:

1. Results of the IR evaluation confirm that transliteration **can not improve** cross-language IR. However, it is not a good strategy to transliterate names that are already translated in the dictionary. **(F)**

Results of the IR evaluation confirm that transliteration **can improve** cross-language IR. However, it is not a good strategy to transliterate names that are already translated in the dictionary.

2. The TRANSLITERATION model is a set of conditional probability distributions over Arabic characters and NULL, conditioned on English unigrams and selected n-grams. **(T)**
3. In the training, they started with a list of 125,000 English proper nouns and their Arabic translations from Arabic Proper Names Dictionary NMSU. **(T)**