Modelling of Terms Across Scripts through Autoencoders

Parth Gupta
Natural Language Engineering Lab
PRHLT Research Center
Universitat Politècnica de València, Spain
pgupta@dsic.upv.es

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

mixed-script information retrieval, term modelling, deep-learning

ABSTRACT

For many languages that use non-Roman based indigenous scripts (e.g., Arabic, Greek and Indic languages) one can often find a large amount of user generated transliterated content on the Web in the Roman script. Such content creates a monolingual or cross-lingual space with more than one scripts which is referred as *mixed-script space* and information retrieval in this space is referred as *mixed-script information retrieval* (MSIR) [1]. In mixed-script space, the documents and queries may either be in the native script and/or the Roman transliterated script for a language (mono-lingual scenario). There can be further extension of MSIR such as multi-lingual MSIR in which terms can be in multiple scripts in multiple languages. Since there are no standard ways of spelling a word in a non-native script, transliteration content almost always features extensive spelling variations. This phenomenon presents a non-trivial term matching problem for search engines to match the native-script or Roman-transliterated query with the documents in multiple scripts taking into account the spelling variations. This problem, although prevalent in Web search for users of many languages around the world, has received very little attention till date. Very recently we have formally defined the problem of MSIR and presented the quantitative study on it through Bing query log analysis [1].

The term-equivalents in the mixed-script space often preserve the sound or pronunciation in the non-native script and the phonemes, captured by the character n-grams, are very important features [2, 3]. The term equivalents across the scripts, including the spelling variants, can be seen as different views for the same term. Deep-learning based autoencoders have shown to perform superior in modelling multi-view (multi-modal) data [4]. In this PhD project we aim to study the modelling of such terms to aid MSIR and more complex problems such as multi-lingual MSIR. Concretely, major questions under this research are: 1) what is the best way to model the terms in mixed-script space? 2) how the retrieval performance is affected by explicit handling of transliterated queries? 3) query formulation and term-weighting for MSIR; 4) what are the properties of different cross-view autoencoder architecture and their impact on retrieval performance? and 5) can the character-level cross-view framework for modeling terms be extended to term-level framework for modelling cross-lingual documents?

In our preliminary research we have have tried to study the distribution of phonemes in the terms and found terms can be better modelled under Dirichlet-multinomial distribution. We have also proposed and investigated the autoencoder based joint model to find term equivalents. The experiments are carried on standard benchmark dataset distributed under FIRE 2013 shared task on Transliterated Search. The experiments suggest that explicit handling of terms for MSIR have strong effect on retrieval and the proposed method shows statistical significant improvement over various state-of-the-art baselines (12% increase in MRR and 29% increase in MAP). More details can be found in [1].

The strong performance of autoencoder over the baselines based on linear dimensionality reduction techniques such as latent semantic indexing (LSI) and canonical correlation analysis (CCA) for MSIR in [1] suggests that the objective function for modelling terms under mixed-script space is non-linear and not convex making autoencoder a proper choice to handle the problem. In future we will carry the experiments based on the questions listed above.

1. REFERENCES

[1] P. Gupta, K. Bali, R. E. Banchs, M. Choudhury, and P. Rosso. Query expansion for mixed-script information retrieval. In *Proceedings of SIGIR*, Gold Coast, Australia, 2014.

[2] K. Knight and J. Graehl. Machine transliteration. *Comput. Linguist.*, 24(4):599–612, Dec. 1998.

[3] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. In *Proceedings of IJCAI*, pages 1360–1365, Barcelona, Spain, July 2011.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of ICML*, pages 689–696, Bellevue, USA, June 2011.

SIGIR'14, Jul 06-11 2014, Gold Coast , QLD, Australia
ACM 978-1-4503-2257-7/14/07.
http://dx.doi.org/10.1145/2600428.2610380.