Cross-lingual clustering

Overview

News articles provided in different languages are clustered when covering the same events. For this purpose a bilingual probabilistic topic model is trained on a comparable multilingual corpus (e.g., Wikipedia) generating interlingual topic representations. The interlingual topic representations are then inferred for the news articles. A clustering is performed based on the interlingual topic representations. The method is completely unsupervised and requires no translation dictionaries.

In depth description

Event detection and clustering are important tasks as a way to gather specific information from large sources, but can also serve as a preliminary step, e.g., for multilingual document summarization. In an effort to eliminate the human effort of creating dictionaries, and to design a system that can perform event detection for any pair of languages, this work proposes using probabilistic topic models that can be trained on freely available comparable corpora. As an example, linked articles from two online Wikipedia encyclopaedias are gathered. As an evaluation, the multilingual Bilingual Latent Dirichlet Allocation method is compared with a vector space model. Whereas the former represents documents from both languages in the same topic space, the vector space model needs translations to find matching terms. A translation dictionary is used to provide these matches. In our evaluations, the probabilistic topic model proves to outperform the translation based vector space model in the multilingual event detection task. The contributions regard the successful application of Bilingual LDA in the real-life application of multilingual event detection.

Potential fields of application

This technology can be included as a part of a media navigation engine that allows visualization of the content of multilingual documents.

Possibilities for exploitation

They include the usage of the component in a national project on cross-lingual mining and search (SCATE, SBO-IWT-130041).

Further Information

Further technical information is available in Vulic, I., De Smet, W., Tang, J. & Moens, M.-F. Probabilistic Topic Modeling in Multilingual Settings: An Overview of Its Methodology and Applications. *Information Processing & Management* (accepted with minor revisions).

Contact person

Prof. Marie-Francine Moens
Department of Computer Science
Celestijnenlaan 200A
B-3001 Heverlee, BELGIUM
sien.moens@cs.kuleuven.be