Learning language-independent sentence representations for multi-lingual, multi-document summarization

Georgios Balikas*†[1] and Massih-Reza Amini*[2]

[1]Laboratoire d'Informatique de Grenoble (LIG) – CNRS : UMR5217, Université Pierre-Mendès-France - Grenoble II, Institut polytechnique de Grenoble (Grenoble INP), Université Joseph Fourier - Grenoble I – UMR 5217 - Laboratoire LIG - 38041 Grenoble cedex 9 - France Tél. : +33 (0)4 76 51 43 61 - Fax : +33 (0)4 76 51 49 85, France

[2]Laboratoire d'Informatique de Grenoble (LIG) – CNRS : UMR5217, Université Pierre-Mendès-France - Grenoble II, Institut polytechnique de Grenoble (Grenoble INP), Université Joseph Fourier - Grenoble I, AMA – UMR 5217 - Laboratoire LIG - 38041 Grenoble cedex 9 - France Tél. : +33 (0)4 76 51 43 61 - Fax : +33 (0)4 76 51 49 85, France

Résumé

This paper presents an extension of a denoising auto-encoder to learn language-independent representations of parallel multilingual sentences. Each sentence from one language is represented using language dependent distributed representations. The input of the auto-encoder is then constituted of a concatenation of the distributed representations corresponding to the vector representations of translations of the same sentence in different languages.
We show the effectiveness of the learned representation for extractive multi-document summarization, using a simple cosine measure that estimates the similarity between vectors of sentences found by the auto-encoder and the vector representation of a generic query represented in the same learned space. The top ranked sentences are then selected to generate the summary. Compared to other classical sentence representations, we demonstrate the effectiveness of our approach on the TAC 2011 MultiLing collection and show that learning language-independent representations of sentences that are translations one from another helps to significantly improve performance with respect to Rouge-SU4 measure.

Mots-Clés: Summarization, Multiview Learning, Machine Learning, Natural Language Processing

*Intervenant

†Auteur correspondant: georgios.balikas@imag.fr