http://lib.uliege.be https://matheo.uliege.be

Master thesis : Term extraction from domain specific texts

Auteur : Poumay, Judicaël
Promoteur(s) : Ittoo, Ashwin
Faculté : Faculté des Sciences appliquées
Diplôme : Master en science des données, à finalité spécialisée
Année académique : 2018-2019
URI/URL : http://hdl.handle.net/2268.2/7487

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.

Term extraction from domain specific texts

In the thesis, we developed a novel unsupervised algorithm for terminology extraction (TE). TE consists in detecting and ranking possible terms from a given document. While a term is a sequence of words that refers to a particular concept in a given domain.

This thesis also brings with it two other ancillary contributions. A new relevancy measure for term ranking; which uses a mix of a termhood, a unithood, and a noise measure to provide a reliable score. And an abbreviation extractor which discovers and extracts the extended form of abbreviated terms using a simple heuristic.

In our review of the literature, we discussed the three main paradigms for TE : linguistic, statistic, and hybrid. While most TE algorithm are hybrid, we noticed the limitations of linguistical technique. Primarily, they are not capable of extracting long and complex terminology. For the statistical approach we discussed the various possible measures that exist to rank terms and that they are divided into two types: termhood and unithood.

As noted, many algorithms already exist for extracting terms but they have limitations. Primarily, we found that no current method was capable of reliably extracting long and complex terminology. Therefore, the algorithm we proposed was designed to handle such task. The extraction of long terms is carried out with the overcomposition of smaller terms.

The results we present from our experiments shows state-of-the-art performance compared to the reviewed literature. More precisely, in our experiments, we evaluated a traditional linguistical approach: POS pattern detection. We demonstrated it was eminently capable of extracting simple terminology but not complex one. We also compared many measures of relevancy and showed that our proposed measure outperformed them all.

Author Judicael POUMAY

supervisor Ashwin ITTOO

Academic year 2018-2019