Research engineer position at Aix Marseille University (France) - SELEXINI project
Data and software support for next-generation word sense induction models

* Contract duration: 12 months (flexible)
* Starting date: June 2022 (flexible)
* Application deadline: May 2, 2022 by email to carlos.ramisch [AT] lis-lab.fr
* Location: LIS lab, TALEP team, Aix Marseille University, Luminy campus, Marseille
* Remuneration (CDD): 1,600€ to 2,000€, depending on experience

SELEXINI is a research project whose goal is to develop next-generation **word sense induction** methods for French. The induced lexicons will not only cluster word usages according to their senses, but also contain multiword expressions, argumental structure, generated definitions, etc. The developed word sense induction methods will build upon large pre-trained language models (e.g. FlauBERT, CamemBERT) and existing lexical resources (French Wiktionary).

We are currently looking for an **engineer whose mission will be to put in place the initial infra-structure of the project**. This mission will have five phases:

(1) **corpus curation:** gather, pre-process and format a large French corpus from diverse written registers;
(2) **preprocessing:** apply in-house tools to (deep) parse, identify multiword expressions, etc. on the corpus;
(3) **pre-lexicon extraction:** extract and structure information from French Wiktionary (e.g. using DBNary);
(4) **language model adaptation**: additional tuning of pre-trained language models on target corpora;
(5) **alignment:** link corpus occurrences to Wiktionary entries, both for words and multiword expressions.

Dealing with such a billion-word corpus will require developing strategies to speed up processing via parallel processing, e.g. to apply pre-processing tools to the corpus. Therefore, the recruited person will have access to the LIS cluster and to dedicated processing nodes. Optimising storage for this great mass of pre-processed text is also a secondary requirement. Finally, the recruited engineer is expected to study and propose the development and/or adaptation of a friendly corpus query platform for the project researchers. This interface will facilitate extracting usages for given lexical units, both for manual inspection and for automatic clustering algorithms.

In addition to these tasks, it will be possible to carry out more exploratory research in the project, according to the interests of the recruited person.

Environment

This position is funded by the ANR SELEXINI project. The recruited person will participate in project meetings, co-author articles, and interact with other SELEXINI members in French research institutions. The engineer will be supervised by Carlos Ramisch in coordination with SELEXINI partners. The recruited person will become a member of the TALEP team, specialised in computational linguistics. TALEP is a dynamic and international team of LIS, a computer science lab located on the Luminy campus in Marseille. Aix Marseille University is one of the largest universities in France, providing a lively and diverse research environment. Depending on the covid-19 situation at the time, partly remote work can be negotiated.

Profile

* Master or PhD in computer science, computer engineering or computational linguistics
* Knowledge of French and English (not necessarily native)
* Interest in languages and familiarity with language technology/NLP

Application

Please send your CV and a few lines explaining why you are applying to carlos.ramisch [AT] lis-lab.fr before May 2.