



Sujet :

Caractérisation d'un espace sémantique bimodal texte-image

Niveau : Bac + 5 - Master

Durée du stage : 4 à 6 mois

Contexte

Les espaces vectoriels sémantiques autrement appelés espaces distributionnels ou espaces de mots (en anglais Word Space Model) permettent, à partir de l'analyse de très grand corpus, de projeter les mots d'une langue dans des espaces vectoriels. Ces espaces sémantiques fournissent ainsi une mesure de proximité entre les mots, plus ou moins proches de ce que l'humain peut interpréter comme étant une proximité sémantique.
Par ailleurs, dans le domaine du traitement d'image, la similarité entre deux images peut se calculer à partir de descripteurs locaux. Ces descripteurs sont des vecteurs calculés sur des points saillants de l'image. Chaque image est constituée d'un certain nombre de descripteurs locaux, de la même façon qu'un texte est constitué d'un certain nombre de mots. On parle ainsi de vocabulaire visuel.

Objectifs du stage

Jusqu'à présent, peu de travaux se sont intéressés à la conception d'espaces multimodaux. Ce stage consiste à explorer cette piste par la constitution d'un espace distributionnel combinant un vocabulaire textuel et un vocabulaire visuel.

Mots textuels et mots visuels seront alors éléments d'un même espace vectoriel.

On suppose que la distance qui sépare les mots visuels des mots textuels dans un tel espace peut avoir du sens pour la perception humaine et éventuellement donner lieu à la création de lexiques bimodaux. Les résultats de ce travail valideront ou non cette hypothèse.

Déroulement du stage

Le stagiaire suivra la démarche suivante :

- Appropriation des outils existants et état de l'art
- Participation à la caractérisation de l'espace bimodal
- Implémentation du générateur de l'espace bimodal
- Tests et évaluation de l'espace sémantique bimodal généré

Bibliographie

Padó, S. & M. Lapata (2007), "Dependency-based construction of semantic space models", Computational Linguistics, 33/2:161-199

Evert, S. & Lenci, A. (2009), Course. <http://wordspace.collocations.de/doku.php/course:start>

Domaines de spécialité requis :

Espaces sémantiques distributionnels, Descripteurs d'images, Traitement Automatique des Langues, Cooccurrences

Moyens informatiques mis en œuvre :

C++ ou Python, environnement Linux



COMMISSARIAT A L'ENERGIE ATOMIQUE
DRT/DTSI/SRCI
Relation avec les écoles : MARC MERGY
Tél : 01 46 54 81 58
E-Mail : marc.mergy@cea.fr

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT - LIST

Dépt/Service/Labo : DTSI/SRCI/LIC2M :

Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue

Web : <http://www-list.cea.fr/>

Encadrement

Nom de l'ingénieur responsable : Gaël De Chalendar Tél. : +33 1 46 54 80 18

E-mail : gael.de-chalendar@cea.fr Fax : +33 1 46 54 75 80

Secrétariat : DTSI/SRCI Tél. : +33 1 46 54 91 17

Nom du chef de laboratoire : Olivier Mesnard Tél. : +33 1 46 54 73 38