HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Hes·so
Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences and Arts Western Switzerland

Cours TAL – Labo 6 : désambiguïsation lexicale (*Word Sense Disambiguation*)

Distribué le mardi 7 mai 2019

Objectif et informations

Comparer deux méthodes de désambiguïsation du sens des mots en contexte (tâche notée WSD), l'une utilisant l'algorithme de Lesk simplifié, et l'autre utilisant word2vec.

Le labo utilisera une ressource contenant 2'369 occurrences du mot anglais « *interest* » annotées chacune avec le sens du mot dans le contexte respectif. Chacune des deux méthodes fonctionne selon le même principe général : comparer le contexte de l'occurrence avec les définitions des sens, et choisir la définition la plus proche. L'algorithme de Lesk définit la proximité comme le nombre de mots en commun, alors que word2vec peut la calculer comme la similarité de deux vecteurs (somme des vecteurs de mots).

Dans une première approche, qui ne nécessite pas de programmation, on vous demande de comparer une occurrence de chaque sens. Dans une seconde approche, qui nécessite un peu de programmation en Python, vous évaluerez les deux systèmes sur l'ensemble des 2'369 occurrences.

Merci d'envoyer votre notebook Jupyter par email au professeur et à l'assistant, avant le **vendredi 17 mai à 23h59**.

Étapes proposées

1. Le fichier de données se trouve à http://www.d.umn.edu/~tpederse/data.html – chercher « *interest* » vers la fin de la page, et prendre le fichier marqué « local copy ». Quel est le format de ce fichier et comment sont annotés les sens ?
2. Bien lire le fichier README associé. Quelles sont les définitions des six sens de « *interest* » annotées dans les données ? De quel dictionnaire viennent-elles ? Où se trouve-t-il en ligne ?
3. Consulter WordNet en ligne et identifier les définitions correspondant aux six sens annotés dans les données. En les combinant éventuellement avec les résultats du (2), écrivez une liste de mots « pleins » (i.e. sans *stopwords*) pour chaque définition.
4. Considérez la première occurrence de « *interest* » dans les données, et notez les mots « pleins » qui entourent « *interest* » (p.ex. toute la phrase). Combien de mots y'a-t-il en commun avec chacune des six définitions ? Quel est donc le sens le plus probable ?

5. En réutilisant le modèle de word2vec entraîné sur Google News fourni par Gensim, calculer les similarités (cosinus) entre les mots du contexte de la première occurrence (somme des vecteurs de chaque mot) et chacune des six définitions. Quel est donc le sens le plus probable ?

    - Suggestion : chercher dans la documentation word2vec comment obtenir très facilement la similarité entre deux ensembles de mots.

6. *Approche manuelle* : appliquer la procédure précédente à une occurrence de « *interest* » pour chaque sens possible (six sens, donc six mots). Combien sur les six sont correctement désambiguïsés par l'algorithme de Lesk, et combien par la méthode word2vec ?

7. *Approche automatique* : implémenter un court programme qui applique la méthode décrite ci-dessus à chacune des 2'369 occurrences de « *interest* » dans le corpus du test.

    - Il faut donc : extraire les mots voisins de chaque occurrence, compter le nombre de mots en commun avec chaque définition, et choisir le sens qui maximise ce nombre (Lesk) ; également, calculer la similarité word2vec entre le contexte et chaque définition, et choisir le sens qui la maximise.

    - Quelles sont les proportions respectives de bonnes réponses de Lesk et de word2vec ? Pouvez-vous les calculer également pour chaque catégorie ?

8. Comment se comparent vos scores avec ceux publiés dans l'article de Pedersen (2000) disponible ici : http://www.aclweb.org/anthology/A00-2009 ?