ANR COMPO

WP3: Memory biases for segmentation

WP3 : Memory biases for segmentation

- Etude des modèles cognitifs pour la segmentation (acquisition du lexique).
- Transfert des principes sous jacents aux modèles cognitifs vers les modèles neuronaux de TAL.

Découverte des unités lexicales d'une langue

- Etant donné une séquence T de lettres ou de phonèmes ou d'unités acoustiques, on souhaite construire une liste d'unité lexicales pertinentes, le lexique.
- Etant donné une séquence X et un lexique L, on souhaite segmenter X selon les unités de L.
- Problème étudié de plusieurs points de vues :
  - Psycho-linguistique
    - Acquisition du lexique chez l'enfant.
  - Linguistique
    - Morphologie : recherches des unités minimales porteuses de sens
    - Lexicologie (expressions polylexicales)
  - Traitement Automatique des Langues
    - Découpage d'un texte en unités d'entrées d'un modèle de TAL
    - Maximisation de la couverture : problème des mots hors vocabulaire

Quels principes de segmentation ?

- Règles typographiques
- Algorithmes de compression : maximisation de la compression
  - Byte Pair Encoding Gage, P. (1994). A new algorithm for data compression. C Users Journal, 12(2), 23-38.
  - WordPiece Schuster, Mike, and Kaisuke Nakajima. "Japanese and korean voice search." ICASSP, 2012.

Quels principes de segmentation ?

- Modèles probabilistes : maximisation de la vraisemblance
  - Goldwater, Sharon, Thomas L. Griffiths, and Mark Johnson. "A Bayesian framework for word segmentation : Exploring the effects of context." Cognition 112.1 (2009) : 21-54.
  - Morfessor Smit, Peter, et al. "Morfessor 2.0 : Toolkit for statistical morphological segmentation." EACL, 2014.
- Modèles cognitifs : détection des régularités statistiques contraintes de mémoire et de traitement
  - Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport. "Statistical learning by 8-month-old infants." Science 274.5294 (1996) : 1926-1928.
  - PARSER Perruchet, Pierre, and Annie Vinter. "PARSER : A model for word segmentation." Journal of memory and language 39.2 (1998) : 246-263.
  - McCauley, Stewart M., and Morten H. Christiansen. "Language learning as language use : A cross-linguistic model of child language development." Psychological review 126.1 (2019)

Passage à l'échelle des modèles cognitifs

- Etude préliminaire : stage de Marianne Schweitzer (MASCO 1)
- Evaluation du modèle PARSER sur un corpus oral (ORFEO)
- Phonétisation du corpus
- Les mots les plus fréquents sont découverts assez vite
- Les scores sont comparables aux fréquences d'occurrences
- La couverture est médiocre
- Le paramètre modélisant l'oubli est difficile à optimiser.

En cours

- Etude du modèle de Goldwater
- Evaluation sur des entrées orthographiques/phonétiques
- Apprentissage long (des dizaines de millers de passage sur les données)

Quelques conclusions intermédiaires

- Les modèles cognitifs ne semblent pas assez mûrs pour proposer des lexiques réalistes pour des modèles de TAL
- Limites des méthodes fondées sur les statistiques
- Beaucoup de modèles, reposant sur des hypothèses différentes
- Scénario : inférer différents lexiques à partir des différents modèles sur les mêmes données puis comparer les lexiques produits

Quels critères pour la comparaison ?

- Perplexité de modèles de langage fondés sur les différents lexiques
- Statistiques des lexiques produits (taille, distribution de la longueur des unités ...)
- Qualité et nature linguistique des lexiques produits (segmentation en unités linguistiques phonèmes, syllabes, morphèmes, mots, locutions, syntagmes)
- Ressources nécessaires pour la construction du lexique (données, puissance de calcul, mémoire)
- Evaluation sur des tâches de modélisation de la compositionnalité comment combiner les représentations associées aux unités ?
- Lien avec des données comportementales.

Quelques questions / perpectives

- Quelles données (orthographiques, phonétiques, acoustiques) ?
- Morphologie non concaténative, langues sémitiques
  - racine + gabarit $\rightarrow$ lemme
  - $ktb + 1A2i3 \rightarrow katib$
- Unités non contigues
- Lifelong learning : dépasser la dichotomie apprentissage/inférence