Reading session

Marine Wauquier

LLF Morphology Workgroup

June 25th, 2021

Article 2

Hofmann, V., Pierrehumbert, J. & Schütze, H. (2020): DagoBERT: Generating Derivational Morphology with a Pretrained Language Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 16–20

- Exploration des modèles de langue pré-entraînés
  - Quelles connaissances en morphologie dérivationnelle ?
- Tâche de *Derivation Generation*
  - Apport des modèles de langue pré-entraînés par rapport aux autres méthodes (LSTM, etc.)

Dispositif

- Tâche de phrase à trou
  - *This jacket is* {*unwearable*}.
- Tâche de classification d'affixes
  - Quel affixe est le plus probable étant donné une base et un contexte ?
  - Prédiction correcte si le mot prédit est la cible masquée
  - Évaluation moyennant le résultat pour chaque affixe basé sur le rang de la cible masquée

- Tuple
  - Phrase avec cible masquée - *This jacket is* {}.
  - Cible - *unwearable*
  - Base - *wear*
- Extraction depuis Reddit
  - Segmentation de chaque mot du corpus
  - Conservation des mots construits dont les affixes et la base sont présents dans le modèle
  - Extraction de chaque phrase contenant au moins un dérivé de base identifiée (10 à 100 mots)
    - Un suffixe, un préfixe, ou un préfixe et un suffixe
  - Seuil de fréquence de 128, et partition en 7 groupes
  - Partition en sets de développement, d'entraînement et de test (20/60/20)

Identification des dérivés

- Algorithme récursif d'identification des bases sur la base
  - Décomposition récursive en parents
  - Affixe et base présents dans le vocabulaire du modèle
    - 48 préfixes et 44 suffixes
  - 20 259 bases, 413 271 dérivés et 123 809 485 contextes
- Précision de la segmentation (100 dérivés) de $0.960 \pm .074$
  - Valeurs plus élevées pour les préfixes préfixes ($0.990 \pm .027$) que les suffixes ($0.930 \pm .093$)

Préfixation

- Problème de l'annotation des bases pour la préfixation dans le modèle
  - *un* et *allowed* dans le modèle, mais pas *##allowed*
  - Plusieurs méthodes de segmentations

Résultats

- DagoBERT (modèle pré-entraîné avec reprise de la segmentation) le plus performant
- Précision qui augmente avec la fréquence (en cas de recouvrement des sets)
- Performance plus faible dans le cas de la génération simultanée du préfixe et du suffixe

Analyse des erreurs

- Congruence entre les affixes ciblés et prédits
  - *-ity* et *-ness*, *-hood* et *-dom*, *-ify* et *-ize*
- Affixe le plus productif favorisé
- Confusion pour certaines relations sémantiques
  - *pro-* et *anti-*, *pre-* et *post-*, *over-* et *under-*