Approche computationnelle de la transparence sémantique

Marine Wauquier

08 octobre 2021

Morphology Reading Group

Contexte

- La mesure de la transparence sémantique est au cœur d'un nombre croissant de travaux
  - Dans une optique de comparaison
    - Bonami & Paperno (2018), Gagné *et al.* (2017), Varvara *et al.* (2021), inter alia
  - Dans une optique de description
    - Bonami & Tribout (2021), Lombard *et al.* (2021), inter alia
- Transparence sémantique comme un proxy pour la notion de lexicalisation, notamment, et pour l'identification de la différence entre sens construit et sens attesté
  - Plus ces sens divergent, moins le mot construit est transparent sémantiquement

Contexte

- Comment mesurer de façon fiable et efficace la transparence sémantique ?
- Une majeure partie de ces travaux repose sur la sémantique distributionnelle
  - Mesure corpus-based de la similarité sémantique entre deux items
    - Gagné et al, Varvara et al, Lombard et al
  - Étude de la variation entre paires d'items
    - Bonami & Paperno, Guzman Naranjo & Bonami, Bonami & Tribout
- D'autres mesures sont-elles envisageables et/ou pertinentes ?

Articles

- SemEval a (indirectement) consacré 2 tâches à cette question

SemEval 2016 Task 11 - Complex Word Identification
Paetzold, G., & Specia, L. (2016). Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 560-569).

SemEval 2021 Task 1 - Lexical Complexity Prediction
Shardlow, M., Evans, R., Paetzold, G. H., & Zampieri, M. (2021). Semeval-2021 task 1: Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 1–16)

Complexité et transparence

- Définition trouble de la complexité
  - SemEval 2016 définit la CWI comme une tâche de décision visant la simplification de certains mots
  - SemEval 2021 définit la LCP comme une tâche d'identification des mots trop difficiles pour le lecteur
    - La complexité y est notamment évaluée en termes de familiarité
- Recouvre la notion de mot inconnu et de sens non prédictible
  - "The occurrence of an unknown word in a sentence can adversely affect its comprehension by readers. Either they give up, misinterpret, or plough on without understanding." (Shardlow et al 2021)
  - Lien avec la notion de transparence, notamment dans le cas des mots construits morphologiquement

Description de la CWI

- **Consigne :** Les participants devaient créer des systèmes permettant, pour un mot cible dans une phrase donnée, prédire si des locuteurs anglophones non natifs pouvaient comprendre le sens du mot cible
  - Tâche binaire (complexe/simple)
  - Évaluation à l'aide de l'Accuracy et du rappel
- 21 équipes ont soumis un total de 42 systèmes

Données pour la CWI

- 9200 phrases annotées manuellement
- Principalement extraites de la version Simple English de Wikipedia
  - Avec différentes contraintes sur les opérations de simplification
- Annotation par 400 volontaires anglophones non natifs
  - Tâche binaire
    - Comprennent-ils le sens du mot cible en contexte ?
    - Comprennent-ils le sens du mot cible hors contexte ?
  - 200 phrases ont été annotées par 20 annotateurs, et les 9000 autres phrases ont été annotées par 1 annotateur
    - Accord faible de 0.244 (pour les 200 phrases)
  - Deux jeux de données d'entraînement (sur les 200 phrases)
    - **joint** - 1 unique label (1 si jugé complexe par au moins 1 annotateur, sinon 0)
    - **decomposed** - 20 label (1 par annotateur)

Description de la LCP

- **Consigne :**
  - Sous-tâche 1 - Assigner une valeur de complexité (de 0 à 1) à toutes les unités monolexicales du corpus
  - Sous-tâche 2 - Assigner une valeur de complexité (de 0 à 1) à toutes les unités - monolexicales ou polylexicales - du corpus
- 55 (ou 56 ?) participations au total

Données pour la LCP

- Total de 5 617 unités annotées dans 10 800 phrases
  - 4 129 unités monolexicales
  - 1 488 unités polylexicales
- Extraction à partir de 3 corpus
  - Europarl (proceedings du Parlement européen)
  - Version anglaise de la Bible
  - Littérature biomédicale issue du corpus CRAFT

Annotation des données

- Mot soumis avec son contexte (et apparaissant dans plusieurs contextes)
- Echelle continue allant de 0 à 5
  - Très facile (0) : mot jugé familier
  - Facile (0.25) : mot dont l'annotateur connaît le sens
  - Neutre (0.5) : mot qui n'est ni difficile, ni simple
  - Difficile (0.75) : mot dont le sens n'est pas limpide pour l'annotateur, mais que l'on peut comprendre en contexte
  - Très difficile (1) : mot que l'annotateur n'a jamais vu ou qu'il ne comprend pas
- Score finale comme la moyenne des scores attribués par les annotateurs
  - En moyenne 17 annotateurs
  - Pas d'information sur l'accord inter-annotateur

Techniquement

- Systèmes qui incluent une variété de méthodes
  - Arbres de décision
  - Classifieurs (SVM...)
  - Régressions
  - Word embeddings et embeddings contextuels
  - Modèles de langue

Features

- Des features variables
  - Fréquence
  - Nombre de syllabes
  - Nombre de synonymes, d'hypéronymes ou d'hyponymes
  - Présence dans des ressources lexicales
  - Entités nommées
  - Informations syntaxiques (catégorie grammaticale)
  - Âge d'acquisition

Et la morphologie ?

- Pas d'information quant à la construction morphologique des mots soumis dans les deux tâches
- Peu de systèmes intégrant des informations morphologiques
  - Environ 6 en 2016, contre 1 en 2021

Exemples

- Stodden, R., & Venugopal, G. (2021, August). RS-GV at SemEval-2021 Task 1: Sense Relative Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 640-649).
- 11 traits morphologiques
  - nom propre, singulier, pluriel
  - taille de la famille, nombre et taille des préfixes et suffixes
- Hypothèse que la complexité (lexicale) d'un mot sera positivement corrélée à sa complexité morphologique

Références

- Bonami, Olivier & Delphine Tribout. 2021. Echantinom: a hand-annotated morphological lexicon of French nouns. DeriMo.
- Gagné, Christina L., Thomas L. Spalding & Kelly A. Nisbet. 2017. Processing English compounds:Investigating semantic transparency. *SKASE Journal of Theoretical Linguistics* 13(2). 2–22.
- Varvara, Rossella, Gabriella Lapesa & Sebastian Padó. 2021. Grounding semantic transparency incontext. *Morphology* 213–234
- https://github.com/MMU-TDMLab/CompLex