Vers les Sens et Au-delà : Induire des Concepts Sémantiques Avec des Modèles de Langue Contextuels

Bastien Liétard Pascal Denis Mikaela Keller
University of Lille, Inria, CNRS, Centrale Lille,
UMR 9189 - CRIStAL, F-59000 Lille, France
prenom.nom@inria.fr

RÉSUMÉ

La polysémie et la synonymie sont deux facettes cruciales et interdépendantes de l'ambiguïté lexico-sémantique, mais elles sont souvent considérées indépendamment dans les problèmes pratiques en TAL. Dans cet article, nous introduisons l'induction de concepts, une tâche non-supervisée consistant à apprendre un partitionnement diffus de mots définissant un ensemble de concepts directement à partir de données. Cette tâche généralise l'induction du sens des mots (via l'appartenance d'un mot à de multiples groupes). Nous proposons une approche à deux niveaux pour l'induction de concepts, avec une vue centrée sur les lemmes et une vue globale du lexique. Nous évaluons le regroupement obtenu sur les données annotées de SemCor et obtenons de bonnes performances (BCubed-F1 supérieur à 0,60). Nous constatons que les deux niveaux sont mutuellement bénéfiques pour induire les concepts et les sens. Enfin, nous créons des plongements dits « statiques » représentant nos concepts induits et obtenons des performances compétitives par rapport à l'état de l'art en Word-in-Context.

ABSTRACT

To Word Senses and Beyond : Inducing Concepts with Contextualized Language Models

Polysemy and synonymy are two crucial interrelated facets of lexical ambiguity and while they have been studied extensively in NLP, they are often being considered independently in practictal problems. In this paper, we introduce Concept Induction, the unsupervised task of learning a soft clustering among words that defines a set of concepts directly from data. This task generalizes Word Sense Induction. We propose a bi-level approach to Concept Induction that leverages both a local lemma-centric view and a global cross-lexicon view to induce concepts. We evaluate the obtained clustering on SemCor's annotated data and obtain good performance (BCubed F1 above 0.60). We find that the local and the global levels are mutually beneficial to induce concepts and also senses. Finally, we create static embeddings representing our induced concepts and obtain competitive performance with the State-of-the-Art on the Word-in-Context task.

MOTS-CLÉS : Sémantique Lexicale, Induction de Sens, Synonymie, Polysémie.

KEYWORDS: Lexical Semantics, Word Sens Induction, Synonymy, Polysemy.

ARTICLE : **Accepté à** 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), voir https://aclanthology.org/2024.emnlp-main.156/.