Université de Lorraine — IDMC
Masters SC & TAL — M1
2019 – 2020

Project Form

Context Similarity and Semantic Relationships

Supervisor:

- Team Sémagramme of the LORIA
- Sylvain Pogodalla, sylvain.pogodalla@inria.fr

Description:

1. **Global Description** In natural language processing (NLP) systems, vector representations of words (word embeddings) are nowadays ubiquitous. These embeddings rely on the distributional hypothesis (Harris 1954): the meaning of a word is provided by the linguistic contexts in which it occurs and semantically related words should be represented by similar vectors.

   However, the exact nature of the semantic relatedness that word and sentence embeddings encode remains unclear. Context similarity mixes distinct relations together (e.g., synonymy, hyponymy, etc. Peirsman, Heylen, and Speelman 2007; Heylen et al. 2008) and it depends on many heuristics and design choices (Padó and Lapata 2007) such as the choice of the similarity measure, the context size, the type of contexts (Weeds, Weir, and McCarthy 2004; Padró et al. 2014).

   On the other hand, some linguistic theories develop models of composition between lexical units, for instance the theory of explanatory combinatorial lexicology, the lexicographical part of the Meaning-Text Theory (Mel'čuk and Polguère 2016), which provides a fine-grained characterization of lexical functions, i.e., the semantic relations between lexical units.

   The goal of the project is to relate word embeddings and context similarity as acquired from textual data to formal theories of semantic relatedness.

2. **Bibliography** (UE 705, semester 7) During this part of the project, the candidate will get familiar with the methods and the theories to be used in the project and with some related work. The candidate will also get familiar with the MANGOES software to be used for experiments.

3. **Implementation** (UE 805, semester 8) During this second part, the candidate will focus (and motivate their choice) on some semantic relations and explore how they are expressed in vector space models, in particular with respect to the various similarity measures, context definition, corpus annotation, from which they are built. The candidate will then propose and implement discovering methods for those semantic relations and evaluate them on different corpora.

Information: This project proposal is part of a larger project including the Sémagramme team in Nancy, the MAGNET team (which develops MANGOES) in Lille, and the DFKI in Saarbrücken.

Deliverables and Schedule: First part of the project:

- October– November: Presentation on distributional semantics.
- November–December: Presentation on the theory of explanatory combinatorial lexicology

and of work relating the latter to distributional.

Second part of the project:

- January–March: Experiments with different input data
- March–April: Proposition of a discovering method, implementation.
- April–May: Analysis of the results; report writing.

References

Harris, Zellig S. (1954). "Distributional Structure". In: *Word* 10.2-3, pp. 146–162. DOI: 10.1080/00437956.1954.11659520.

Heylen, Kris et al. (May 2008). "Modelling Word Similarity: an Evaluation of Automatic Synonymy Extraction Algorithms." In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/818_paper.pdf.

Mel'čuk, Igor and Alain Polguère (2016). "La définition lexicographique selon la Lexicologie Explicative et Combinatoire". In: *Cahiers de lexicologie* 109, pp. 61–91. DOI: 10.15122/isbn.978-2-406-06861-7.p.0061.

Padó, Sebastian and Mirella Lapata (2007). "Dependency-Based Construction of Semantic Space Models". In: *Computational Linguistics* 33.2, pp. 161–199. DOI: 10.1162/coli.2007.33.2.161. URL: https://www.aclweb.org/anthology/J07-2002.

Padró, Muntsa et al. (2014). "Comparing Similarity Measures for Distributional Thesauri". In: *Proceedings of LREC 2014*. Ed. by Nicoletta Calzolari et al. URL: https://www.aclweb.org/anthology/L14-1496/.

Peirsman, Yves, Kris Heylen, and Dirk Speelman (2007). "Finding semantically related words in Dutch: co-occurrences versus syntactic contexts". In: *Proceedings of the 2007 Workshop on Contextual Information in Semantic Space Models: Beyond Words and Documents*. Ed. by Marco Baroni, Alessandro Lenci, and Magnus Sahlgren, pp. 9–16. URL: https://www.semanticscholar.org/paper/Finding-semantically-related-words-in-Dutch:-versus-Peirsman-Heylen/206a195aa08fdde09623b098b0a15cb716e8cc15.

Weeds, Julie, David Weir, and Diana McCarthy (August 23–27, 2004). "Characterising Measures of Lexical Distributional Similarity". In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland: COLING, pp. 1015–1021. URL: https://www.aclweb.org/anthology/C04-1146.

Université de Lorraine — IDMC
Masters SC & TAL — M1
2019 – 2020

Fiche de projet tutoré

Similarité de contextes et relations sémantiques

Encadrement

- Équipe Sémagramme du LORIA
- Sylvain Pogodalla, sylvain.pogodalla@inria.fr

Description :

1. **Description globale** La grande majorité des systèmes de traitement automatique des langues naturelles (TALN) utilisent désormais des représentations de mots sous forme de vecteurs. Ces plongements s'appuyent sur l'hypothèse distributionnelle (Harris 1954) : le sens d'un mot est fourni par son contexte, et des mots sémantiquement reliés devraient être représentés par des vecteurs similaires.

   Alors que le plongement de mots et de phrases encodent une certaine forme de proximité sémantique, sa nature demeure imprécise. Différentes relations (par exemple la synonymie, l'hyponymie, etc. Peirsman, Heylen et Speelman 2007 ; Heylen et al. 2008) sont exprimées par la similarité de contextes, et cela peut varier également suivant les différentes heuristiques choisies (Padó et Lapata 2007) telles que le choix des mesures de similarité, la taille du contexte ou le type de contexte (Weeds, Weir et McCarthy 2004 ; Padró et al. 2014).

   Par ailleurs, certaines théoties linguistiques développent des modèles de composition entre les unités lexicales, par exemple la théorie de la lexicologie explicative et combinatoire, la partie lexicologique de la théorie sens-texte (Mel'čuk et Polguère 2016), qui caractérise avec précision les fonctions lexicales, c'est-à-dire les relations sémantiques entre unités lexicales.

   Le but du projet est d'étudier la relation entre plongements de mots et similarité de contextes, tels qu'appris à partir de données textuelles, et les théories des relations sémantiques.

2. **Bibliographie** (UE 705, semestre 7) Durant cette partie du projet, la candidate ou le candidat se familiarisera avec les méthodes et les outils utilisés dans le projet, ainsi qu'avec des travaux similaires. La candidate ou le candidat se familiarisera également avec le logiciel MANGOES qui sera utilisé dans les expériences.

3. **Réalisation** (UE 805, semestre 8) Durant cette deuxième partie du projet, la candidate ou le candidat se concentrera sur certaines relations sémantiques et étudiera la manière dont elles s'expriment dans le modèle sémantique vectoriel, en particulier vis-à-vis des mesures de similarité, de définition du contexte et des annotations à partir desquelles elles sont construites. La candidate ou le candidat proposera et implantera alors des méthodes pour reconnaître ces relations sémantiques et les évaluera sur différents corpus.

Informations diverses : Ce projet fait partie d'un projet plus large qui comprend les équipes Sémagramme de Nancy, MAGNET (qui développe MANGOES) de Lille et le DFKI de Sarrebruck.

Livrables et échéancier : Première partie du projet :

- octobre– novembre : présentation sur la sémantique distributionnelle

- novembre–décembre : présentation de la théorie de la lexicologie explicative et combinatoire et des travaux la situant par rapport à la sémantique distributionnelle

Deuxième partie du projet :

- janvier–mars : expériences avec les différents types de données et de mesure
- mars–avril : proposition d'une méthode de reconnaissance et son implantation
- avril–mai : analyse des résultats et écriture du rapport

Références

Harris, Zellig S. (1954). "Distributional Structure". In : *Word* 10.2-3, p. 146-162. DOI : 10.1080/00437956.1954.11659520.

Heylen, Kris et al. (mai 2008). "Modelling Word Similarity : an Evaluation of Automatic Synonymy Extraction Algorithms." In : *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco : European Language Resources Association (ELRA). URL : http://www.lrec-conf.org/proceedings/lrec2008/pdf/818_paper.pdf.

Mel'čuk, Igor et Alain Polguère (2016). "La définition lexicographique selon la Lexicologie Explicative et Combinatoire". In : *Cahiers de lexicologie* 109, p. 61-91. DOI : 10.15122/isbn.978-2-406-06861-7.p.0061.

Padó, Sebastian et Mirella Lapata (2007). "Dependency-Based Construction of Semantic Space Models". In : *Computational Linguistics* 33.2, p. 161-199. DOI : 10.1162/coli.2007.33.2.161. URL : https://www.aclweb.org/anthology/J07-2002.

Padró, Muntsa et al. (2014). "Comparing Similarity Measures for Distributional Thesauri". In : *Proceedings of LREC 2014*. Sous la dir. de Nicoletta Calzolari et al. URL : https://www.aclweb.org/anthology/L14-1496/.

Peirsman, Yves, Kris Heylen et Dirk Speelman (2007). "Finding semantically related words in Dutch : co-occurrences versus syntactic contexts". In : *Proceedings of the 2007 Workshop on Contextual Information in Semantic Space Models : Beyond Words and Documents*. Sous la dir. de Marco Baroni, Alessandro Lenci et Magnus Sahlgren, p. 9-16. URL : https://www.semanticscholar.org/paper/Finding-semantically-related-words-in-Dutch:-versus-Peirsman-Heylen/206a195aa08fdde09623b098b0a15cb716e8cc15.

Weeds, Julie, David Weir et Diana McCarthy (23-27 août 2004). "Characterising Measures of Lexical Distributional Similarity". In : *COLING 2004 : Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland : COLING, p. 1015-1021. URL : https://www.aclweb.org/anthology/C04-1146.