Originalveröffentlichung in: Geierhos, Michaela (Hrsg.): DHd2022: Kulturen des digitalen Gedächtnisses. Konferenzabstracts.
Universität Potsdam & Fachhochschule Potsdam, 07. bis 11. März 2022. - Potsdam: Zenodo, 2022. S. 342-344.
DOI: *https://doi.org/10.5281/zenodo.6304590*

Semantische Suche mit Word Embeddings für ein mehrsprachiges Wörterbuchportal

Tu, Ngoc Duyen Tanja
tu@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Germany

Meyer, Peter
meyer@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Germany

Das Lehnwortportal Deutsch (LWPD) (Meyer/Eppinger 2019; lwp.ids-mannheim.de) ist ein Online-Informationssystem zu Entlehnungen von Wörtern aus dem Deutschen in andere Sprachen. Es beruht auf einer wachsenden Zahl von lexikographischen Ressourcen zu verschiedenen Sprachen und bietet eine einfache ressourcenübergreifende Suchfunktion an. Das Poster präsentiert eine derzeit in Entwicklung befindliche onomasiologische Suchfunktion für das LWPD.

Ähnliche Projekte, z.B. van der Sijs (2015), nutzen für die Implementierung ihrer semantischen Suche eigens für ihre Datenbasis erstellte Taxonomien von semantischen Feldern. Eine sehr komplexe Open-Source-Taxonomie findet sich beispielsweise auf semdom.org/. Solche Klassifikationen ziehen häufig folgende Probleme nach sich: (a) Aufgrund der inhärenten Vagheit von Definitionen für semantische Felder beruht die Zuordnung von Einzelbedeutungen einer lexikalischen Einheit zu Feldern immer auf einer subjektiven Annotationspraxis, die (b) von dem:der Nutzer:in gewissermaßen rekonstruiert werden muss; (c) bezogen auf die Taxonomie ist es schwierig, einen guten Kompromiss zwischen einfacher Handhabung und Detailgenauigkeit zu finden; (d) grundsätzliche Änderungen an der Taxonomie sind mit einem hohen Aufwand verbunden.

Wir haben einen alternativen Ansatz implementiert, um die oben genannten Probleme anzugehen. Die technische Umsetzung unserer Methode basiert auf den ConceptNet NumberBatch Word Embeddings (CN) (Speer/Chin/Havasi 2017), die auf multilingualen Daten sowie semantischen Beziehungen zwischen Wörtern trainiert sind. Ein im Grunde ähnlicher Lösungsansatz wurde erfolgreich zur Optimierung von Suchmaschinen genutzt (Castro Fernandez et al. 2018; Kuzi/Shtock/Kurland 2016). Da wir keinen Zugriff auf die Korpusdaten haben, die den lexikographischen Ressourcen des LWPD zugrunde liegen, ist es uns nicht möglich, selbst Word Embeddings zu trainieren.

Für die Implementierung der semantischen Suche werden zunächst jedem Wort im LWPD (darunter deutsche Etyma, Lehnwörter, etc.) mindestens ein Wort sowie der/die entsprechende/n Vektor/en aus CN zugeordnet. Für jedes zugeordnete CN-Wort wird angegeben, in welcher semantischen Beziehung (z.B. Synonym, Hyperonym) es zu dem LWPD-Wort steht. Die Zuordnung wird folgendermaßen durchgeführt:

(1) Wenn ein monosemes LWPD-Wort in CN enthalten ist, wird ihm als Default-Wert automatisch dieses CN-Wort zugeordnet.

(2) Wenn ein LWPD-Wort nicht in CN enthalten ist, aber ein LWPD-Wort, das in einer etymologischen oder Derivationsbeziehung zu ihm steht, dann wird ihm als Default-Wert dieses CN-Wort zugeordnet.

(3) Wenn ein LWPD-Wort polysem ist, wird jeder Bedeutung manuell ein CN-Wort zugeordnet.

(4) Wenn ein LWPD-Wort nicht in CN enthalten ist, wird ihm manuell ein semantisch ähnliches CN-Wort zugeordnet.

Ebenfalls können die Default-Werte manuell geändert werden. Homonymen LWPD-Wörtern werden manuell eindeutige CN-Wörter zugeordnet.

Einem LWPD-Wort bzw. einer Bedeutung eines Wortes können auch mehrere CN-Wörter zugeordnet werden, u.a. um Word Embeddings polysemer Wörter zu disambiguieren. Es wird dann eine gewichtete und normierte Summe der Vektoren der einzelnen zugeordneten CN-Wörter zugrunde gelegt. Das Gewicht eines CN-Wortes ergibt sich aus der angegebenen semantischen Beziehung, die auf eine Ganzzahl abgebildet wird.

Die semantische Suche im LWPD läuft aus Nutzer:innensicht folgendermaßen ab: Es steht eine große Anzahl an häufig verwendeten deutschen Wörtern (im Folgenden: Suchschlüssel) zur Auswahl, die mit automatischer Vervollständigung eingegeben werden können. Mit diesen kann der:die Nutzer:in beliebige Aspekte lexikalischer Bedeutung beschreiben. Alle Suchschlüssel sind in CN enthalten. Somit berechnet sich die semantische Ähnlichkeit der Suchschlüssel und der, den Bedeutungen der LWPD-Wörter zugeordneten, CN-Wörter aus ihrer Kosinus-Ähnlichkeit. Wenn die Kosinus-Ähnlichkeit über einem bestimmten Schwellenwert liegt, werden die entsprechenden LWPD-Wörter in der Suchergebnisliste angezeigt. Die Kosinus-Ähnlichkeiten zwischen den Suchschlüsseln und den CN-Wörtern liegen vorberechnet im LWPD vor.

Um die Qualität der Suchergebnisse zu evaluieren, wurde eine Vorstudie durchgeführt:

(1) Die Bedeutungsangabe (z.B. für das Etymon *Riss*: „ *Spalte, Einschnitt, Einriss*“) von jedem Etymon aus dem LWPD wird lemmatisiert und POS-getagged.

(2) Für jedes Etymon E und jedes Lemma L aus einer zugehörigen Bedeutungsangabe wird mit Hilfe von GermaNet (Hamp/Feldweg 1997; Henrich/Hinrichs 2010) ihre semantische Ähnlichkeit gemäß dem Maß in Lin (1998) ermittelt. Budanitsky/Hirst (2006) zeigen, dass die von menschlichen Annotatoren vergebenen semantischen Ähnlichkeitsscores bei englischen Wortpaaren mit den berechneten Scores des Maßes in Lin (1998) eine hohe Korrelation aufweisen. Wir unterstellen, dass Etyma mit den Lemmata ihrer Bedeutungsangaben typischerweise in einer engen semantischen Relation stehen. Daher werden jeweils die Lin-ähnlichsten Synsets von E und L zugrunde gelegt.

(3) Die Kosinus-Ähnlichkeit zwischen dem Vektor von E (z.B. *Riss*) und dem von L (z.B. *Spalte*) wird berechnet.

(4) Die Kosinus-Ähnlichkeit wird mit dem Ergebnis des Ähnlichkeitsmaßes in (2) verglichen.

Es ergibt sich, dass in unserem LWPD-Datensatz die Kosinus-Ähnlichkeit mit dem Ähnlichkeitsmaß von Lin (1998) positiv korreliert ist (r=0,52) und für Lin-Ähnlichkeit größer als 0,9 auf fast 0,65 ansteigt.

Diese Evaluation gibt allerdings nur einen ersten Hinweis dazu, dass unser Ansatz vielversprechend ist. Sobald die hier präsentierte semantische Suche implementiert und im LWPD verfügbar ist, soll anhand einer Benutzungsstudie die Qualität der Suchergebnisse untersucht werden.

Creative Commons - Attribution 4.0 International

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-114539

Bibliographie

Budanitsky, Alexander / Hirst, Graeme (2006): "Evaluating WordNet-based Measures of Lexical Semantic Relatedness", in: *Computational Linguistics* 32(1): 13-47.

Castro Fernandez, Raul / Mansour, Essam / Qahtan, Abdulhakim A. / Elmagarmid, Ahmed / Ilyas, Ihab / Madden, Samuel / Ouzzani, Mouhrad / Stonebraker, Michael / Tang, Nan (2018): "Seeping Semantics: Linking Datasets Using Word Embeddings for Data Discovery ", in: *Proceedings of the 34th International Conference on Data Engineering* 989-1000.

Hamp, Birgit / Feldweg, Helmut (1997): "GermaNet - a Lexical-Semantic Net for German ", in: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* 9-15.

Henrich, Verena / Hinrichs, Erhard (2010): "GernEdiT - The GermaNet Editing Tool ", in: *Proceedings of the Seventh Conference on International Language Resources and Evaluation* 2228–2235.

Kuzi, Saar / Shtok, Anna / Kurland, Oren (2016): "Query Expansion Using Word Embeddings", in: *Proceedings of the 25th ACM International Conference on Information and Knowledge Management* 1929–1932.

Lin, Dekang (1998): "An Information-Theoretic Definition of Similarity", in: *Proceedings of the Fifteenth International Conference on Machine Learning* 296–304.

Meyer, Peter (2019): "Leistungsfähige und einfache Suchen in lexikografischen Datennetzen. Ein Query Builder für lexikografische Property-Graphen", in: *Digital Humanities: multimedial & multimodal. 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. Konferenzabstracts* 312-314.

Meyer, Peter / Eppinger, Mirjam (2018): "fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data", in: *Proceedings of the XVIII EURALEX International Congress* 1017-1022.

Speer, Robyn / Chin, Joshua / Havasi, Catherine (2018): "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge", in: *arXiv:1612.03975 [cs]*.

van der Sijs, Nicoline (2015): Uitleenwoordenbank, uitleenwoordenbank.ivdnt.org, hosted by the Instituut voor de Nederlandse Taal. Accessed at: http://uitleenwoordenbank.ivdnt.org/. (06 April 2021)