Belegvariation

Ulf Hamster
Berlin-Brandenburgische Akademie der Wissenschaften

Inhalt

Problemanalyse

Lösungsansatz

Übungen mit Colab

Fallbeispiele (nächste Woche)

Problemanalyse

Beobachtung

- Suchtreffer weisen *Dubletten* auf
- … beziehen sich auf die gleiche Lesart, Ereignis, linguistisches Konzept (*Semantik*)
- … haben ähnlichen Satzbau (*Grammatik*)
- … stammen aus/von denselben *Quellen* (Autor/innen, Werke, Zeiträume)

Ursachen

- Suchtreffer mit einem *gleich hohen Score* können sich auf die gleiche Menge von Bewertungsmerkmalen beziehen.
- Scoringmodelle bewerten einzelne Satzbelege *unabhängig voneinander.*

Beispiel: Dubletten für Suchbegriff “Regierungsauftrag”

Exakte *Dubletten*
in mehreren Zeitungsmanteln

Unterschiedliche *Quellen*
bevorzugen, ist nicht genug.

Andere Beispiele:
- Clickbait Repostings,
- Errata in Webartikeln,
- Neuauflage von Büchern
- Übernahme von Zitaten
- Retweets

1: Norddeutsche Neueste Nachrichten, 16.09.2022
Obwohl die Moderaten mit 19,1 Prozent weniger Stimmen erhielten als die Rechten, übernehmen sie den **Regierungsauftrag**.

2: Neue Osnabrücker Zeitung, 16.09.2022
Obwohl die Moderaten mit 19,1 Prozent weniger Stimmen erhielten als die Rechten, übernehmen sie den **Regierungsauftrag**.

3: Der Prignitzer, 16.09.2022
Obwohl die Moderaten mit 19,1 Prozent weniger Stimmen erhielten als die Rechten, übernehmen sie den **Regierungsauftrag**.

4: Schweriner Volkszeitung, 16.09.2022
Obwohl die Moderaten mit 19,1 Prozent weniger Stimmen erhielten als die Rechten, übernehmen sie den **Regierungsauftrag**.

5: Bote der Urschweiz, 09.09.2022
Seit 1955 - ihr erster Premier Winston Churchill war bereits im Amt - hat sie elf Männern und drei Frauen im Buckingham-Palast den **Regierungsauftrag** erteilt und damit ihre wichtigste konstitutionelle Aufgabe erfüllt . (sbo)

6: Luzerner Zeitung, 09.09.2022
Seit 1955 - ihr erster Premier Winston Churchill war bereits im Amt - hat sie elf Männern und drei Frauen im Buckingham-Palast den **Regierungsauftrag** erteilt und damit ihre wichtigste konstitutionelle Aufgabe erfüllt . (sbo)

7: St. Galler Tagblatt, 09.09.2022
Seit 1955 - ihr erster Premier Winston Churchill war bereits im Amt - hat sie elf Männern und drei Frauen im Buckingham-Palast den **Regierungsauftrag** erteilt und damit ihre wichtigste konstitutionelle Aufgabe erfüllt . (sbo)

8: Thurgauer Zeitung, 09.09.2022
Seit 1955 - ihr erster Premier Winston Churchill war bereits im Amt - hat sie elf Männern und drei Frauen im Buckingham-Palast den **Regierungsauftrag** erteilt und damit ihre wichtigste konstitutionelle Aufgabe erfüllt . (sbo)

9: Der Tagesspiegel, 09.09.2022
Seit 1955 - ihr erster Premier Winston Churchill war bereits im Amt - hatte sie elf Männern und drei Frauen stets im Buckingham-Palast den **Regierungsauftrag** erteilt.

10: Frankfurter Rundschau, 07.09.2022
Die nur von ihrer Partei gekürte britische Premierministerin holt sich in Balmoral ihren **Regierungsauftrag** ab / Von Sebastian Borger

Korpustreffer für »Regierungsauftrag«, aus dem Korpus DWDS-Zeitungskorpus (ab 1945) des Digitalen Wörterbuchs der deutschen Sprache, Zeitraum 2021-2022.

Lösungsansatz 1/2 – Das MECE-Prinzip

MECE = "mutually exclusive and collectively exhaustive" (Mengenlehre)

“mutually exclusive” bzw. disjunkt

- Die Menge aller ausgewählten Satzbelege solle keine bzw. *minimale Schnittmengen* hinsichtlich Bedeutungsähnlichkeit und grammatischer Ähnlichkeit, u.a. aufweisen.

“collectively exhaustive”

- Die *Vereinigungsmenge* aller ausgewählten Satzbelege sollte (idealerweise) das gesamte Bedeutungsspektrum umfassen

Lösungsansatz 2/2 – Quadratische Optimierung als Suchfilter

Zum Sortieren der Suchtreffer ermittle Gewichtungen w_i
durch Maximieren der Bewertungsgüte g_i
und Minimieren der aggregierten Ähnlichkeitsmetriken Q_{ij} zw. den Satzbelegen.

$$\min_{w_1,\ldots,w_N} \; -\lambda \cdot \underbrace{\sum_{i=1}^{N} w_i g_i}_{\text{total goodness}} + (1-\lambda) \underbrace{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} w_i Q_{i,j} w_j}}_{\text{total similarity}}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} w_i = 1$$

$$w_i \geq 0 \quad \forall i$$

$$w_i \leq b \quad \forall i$$

$$Q_{i,j} = \sum_{k}^{M} \beta_k D_{i,j}^{(k)} \quad \forall i,j$$

In der Software: ***Nutzer/innen*** können fünf ***Präferenzparameter*** per Slider-UI angeben:

1. Bewertungsgüte vs. Belegvariation (λ)
2. Semantische Ähnlichkeit ‘abstrafen’ (β_1)
3. Satzbau-Ähnlichkeit ‘abstrafen’ (β_2)
4. Dubletten ‘abstrafen’ (β_3)
5. Quellen-Ähnlichkeit ‘abstrafen’ (β_4)

Übungen mit Colab

Übung 1/4

Was bedeutet hier *semantische Ähnlichkeit*?

- Representationsvektoren je Satzbeleg werden mit SBert erzeugt (Contextual Sentence Embeddings) und
- die Cosinus-Ähnlichkeitsmetrik D_ij für ein Paar von Satzbelegen i und j berechnet.

Übung

- Semantische Ähnlichkeiten mit SBert berechnen
- https://colab.research.google.com/drive/1WJXf0g_ty2B_z_-uhMBtnJFVFxKdDHVn

Übung 2/4

Wie werden (Near)-Dubletten und ähnliche Quellenangaben erkannt?

- k-Shingling: Rohtext (inkl. Leerzeichen, Interpunktion) wird in jede mögliche Zeichenketten zerlegt bis max. Zeichenlänge k.
- Beispiel: “abcab” mit k=3 => {a, ab, abc, b, bc, bca, c, ca, cab}
- Ein MinHash aus den Shingles erzeugen
- Berechne Jaccard-Ähnlichkeit zweier MinHashes

Übung

- Fingerprinting mit MinHash/LSH
- https://colab.research.google.com/drive/1_6UUOM7bGFqR1JPInCp5YWPP6WFM3Nb7?usp=sharing

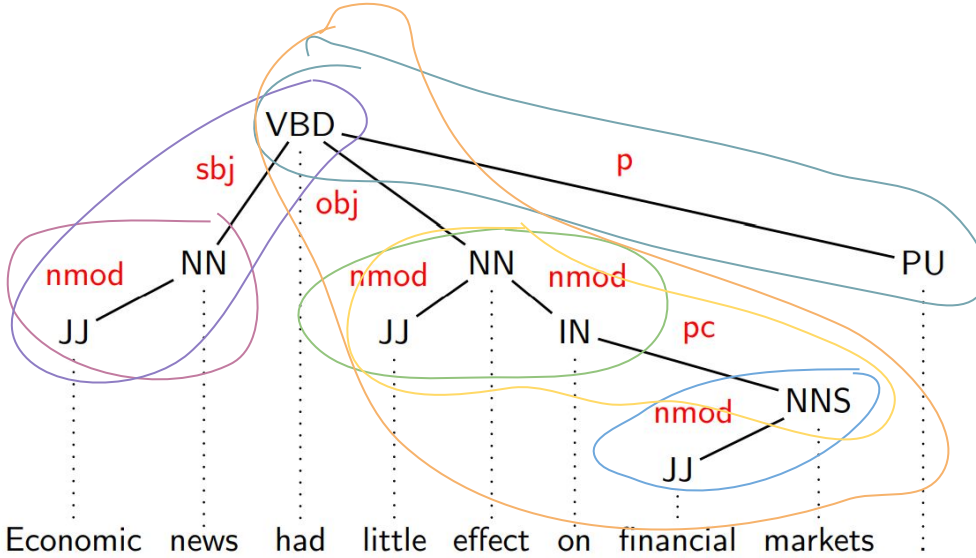
Übung 3/4

Wie wird die syntaktisches Ähnlichkeit ermittelt?

- Dependenzbaum wird in Teilbäume zerlegt (treesimi Paket)
- Serialisiere Teilbäume zu Shingles
- Erzeuge MinHash aus den Shingles eines Satzbeleges
- Berechne Jaccard-Ähnlichkeit zweier MinHashes

Übung

- Syntaktische Ähnlichkeiten berechnen
- https://drive.google.com/file/d/1pmcegCQdW2u1dZUdYUi-_GnE2pElmg21/view?usp=sharing



Übungen 4/4

Quadratische Optimierung mit SBert-Ähnlichkeiten

https://drive.google.com/file/d/1clOkRlBmQ8hyv3Oo7FX1ueluzPSed_Xd/view?usp=sharing

Reverse Automatic Differentiation als Approximator für Quadratische Optimierungsprobleme

$$
\begin{aligned}
\mathcal{L} = & -\lambda \cdot \sum_{i=1}^{N} u_i g_i \\
& + (1-\lambda) \cdot \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} u_i Q_{i,j} u_j} \\
& + \alpha_1 \cdot \left(1 - \sum_{i=1}^{N} w_i\right)^2 \\
& + \alpha_2 \cdot \sum_{i=1}^{N} -\min(0, w_i) \\
& + \alpha_3 \cdot \sum_{i=1}^{N} -\min(0, b - w_i)
\end{aligned}
$$

$$
\begin{aligned}
v_i &= w_i - \min(0, w) \; \forall i \\
u_i &= \frac{v_i}{\max(1e^8, \sum_{j=1}^{N} v_j)}
\end{aligned}
$$

PyTorch und Tensorflow sind “Reverse Automatic Differentiation” Bibliotheken.

Wir müssen das Optimierungsproblem als Loss Function umschreiben.

Die Nebenbedingungen sind dann Reguliarization Penalties.

Quadratische Optimierung mit Tensorflow in Python
https://github.com/satzbeleg/keras-quadopt/blob/main/keras_quadopt/problem.py

Refactored Code in TFJS
https://github.com/satzbeleg/evidence-app/blob/main/src/components/variation/quadopt.js

Fallbeispiele

Käse 1/2 – Nur die Bewertungsgüte (λ=100%)

Top-Suchtreffer:
v.a. Käse als Nahrungsmittel

Außer Brot und Wein wurden noch lange Öl (zur Taufe?), Käse, Oliven, Erstlingsfrüchte (S. 115 ed. Hauler; Klauser- Rech, RAC II, 500), Blumen u. a. »geopfert«, zum Unterhalt von Armen und Klerikern.

w_0: 0.0073 | Schweizer, E. u. a.: Abendmahl. In: Galling, Kurt (Hg.), Die Religion in Geschichte und Gegenwart, Berlin: Directmedia Publ. 2000 [1957], S. 220

(Kaum) drei Käse hoch sein: (noch) ganz klein sein, spöttisch vor allem von einem kleinen Gernegroß gesagt, einem (Drei-)Käsehoch; schon 1767 im 'Versuch eines bremisch-niedersächsischen Wörterbuchs' (Band 2, S. 762):» Een Junge twe Kese hoog: ein kleiner kurzer Junge «; im niederdeutschen Raum machte man früher auf allen Höfen Käse nach Art der (Holländer)

w_1: 0.0073 | Röhrich, Lutz: Käse. In: Lexikon der sprichwörtlichen Redensarten [Elektronische Ressource], Berlin: Directmedia Publ. 2000 [1994], S. 27405

– Käse und Butter, Lapšin.

w_2: 0.0067 | Schlögel, Karl: Petersburg, München Wien: Carl Hanser Verlag 2002, S. 223

Dieser Erstlingskäse ist ein vollfetter Käse von erstklassigem Geschmack und hohem Nährwert.

w_3: 0.0067 | Die Landfrau, 24.01.1925

Die anderen, die vom Jungen beliefert worden waren, gaben uns einen Trostschluck und einen Happen Käse und neckten uns dafür.

w_4: 0.0067 | Alexander Granach, Da geht ein Mensch: Leck: btb Verlag 2007, S. 307

Käse 2/2 – Unterschiedliche Semantik (λ=0%, β_1=100%)

digitalisieren 1/2 – Güte vs. Semantik & Grammatik (λ=25%, β_1=50%, β_2=50%)

Viele Satzbelege aus einer Quelle (Die ZEIT)

Der Protest mit Massen-Mails sollte das Rathaus blockieren, in dem alle Entscheidungsprozesse digitalisiert sind.

w_{13}: 0.0184 | Die Zeit, 29.10.1998, Nr. 45

Die gemeinnützige Stiftung, deren 35-Millionen-Dollar-Haushalt von der amerikanischen Regierung mitfinanziert wird, digitalisiert und katalogisiert die Videobänder zu einer »digitalen Bibliothek«.

w_0: 0.0184 | konkret, 1996

Wird Digital-Fernsehen vor allem mit Bezahlfernsehen identifiziert, so dürfte es schwer werden, das soeben von der Bundesregierung proklamierte Ziel zu erreichen und bis zum Jahr 2010 die Übertragungstechnik komplett zu digitalisieren.

w_1: 0.0184 | Die Zeit, 03.09.1998, Nr. 37

Das Sterben wird also zügig digitalisiert, das Leben nach dem Tod nicht minder.

w_{24}: 0.0183 | Die Zeit, 03.05.1996, Nr. 19

Ähnlich wie bei der Aufnahme einer Compact Disc (CD) wird das analoge Sprachsignal zunächst digitalisiert.

w_{48}: 0.0183 | o.A., Der Computer, der aufs Wort gehorcht, in: Süddeutsche Zeitung 16.02.1995, S. 24

digitalisieren 2/2 – unterschiedl. Quellen (λ=25%, β_1=50%, β_2=50%, **β_4=100%**)

jetzt
ZEIT, C't, konkret,
unbekannt, Bild

Vor allem sollten wir nach den Menschen fragen, die darüber entscheiden, welche Information digitalisiert wird und welche nicht; ob uns in der Flut der Netzinformationen noch Zeit zur Erinnerung und Gedächtnis genug bleibt, uns an anderes zu erinnern als an die perfekte Beherrschung des Netzes und seiner Chancen.

w_3: 0.0188 | Die Zeit, 28.06.1996, Nr. 27

Dabei besteht das geringste Problem darin, alle Bilder zu digitalisieren.

w_{45}: 0.0185 | C't, 2000, Nr. 8

Die gemeinnützige Stiftung, deren 35-Millionen-Dollar-Haushalt von der amerikanischen Regierung mitfinanziert wird, digitalisiert und katalogisiert die Videobänder zu einer »digitalen Bibliothek«.

w_0: 0.0184 | konkret, 1996

Und was wäre, wenn man jedes der schätzungsweise 60 Millionen existierenden Bücher nur einmal digitalisieren würde?

w_{59}: 0.0183 | 37

Die weltberühmte Gutenbergbibel (aus dem Jahr 1456) gibt es jetzt digitalisiert.

w_{21}: 0.0183 | o.A., NEWS-Gute-Schlechte, in: Bild 23.03.2000, S. 4

blau 1/2 – Semantik (λ=0%, β_1=100%)

blau 2/2 – Mehr Variation im Satzbau (λ=0%, β_1=50%, β_2=50%)

Quellen - Literatur

[1] Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P., 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. Presented at the Proceedings of the 13th EURALEX International Congress, pp. 425–432.

[2] Didakowski, J., Lemnitzer, L., Geyken, A., 2012. Automatic example sentence extraction for a contemporary German dictionary, in: Proceedings of the 15th EURALEX International Congress. Presented at the EURALEX 2012, Department of Linguistics and Scandinavian Studies, University of Oslo, Oslo,Norway, pp. 343–349.

Quellen – Software

Softwarerepositorien für das EVIDENCE-Projekt: https://github.com/satzbeleg

Quadratische Optimierung mit Keras: https://github.com/satzbeleg/keras-quadopt

Best-Worst-Scaling Rankings verarbeiten: https://github.com/satzbeleg/bwsample

Contextual Sentence Embeddings mit SBert:
https://github.com/UKPLab/sentence-transformers

Dependenzparsing mit trankit: https://github.com/nlp-uoregon/trankit

Bäume in Teilbäume zerlegen mit treesimi: https://github.com/satzbeleg/treesimi

MinHash/LSH mit Datasketch: https://github.com/ekzhu/datasketch

Strings shinglen: https://github.com/ulf1/kshingle