Softwareprojekt

Katja Markert: Themen SWP 2019

Computerlinguistik
Universität Heidelberg
Sommersemester 2019

Themenvorschläge

Variationen

Alle Themen bieten Raum für mehr als eine Gruppe durch Variation der Methodik, Sprache/Domänen oder Daten. Jede Gruppe kann ihren Fokus (unter Anleitung) selbst bestimmen. Bei gemeinsamen Daten können Gruppen auch in einen "Wettbewerb" miteinander treten (siehe Semeval oder Conll competitions).

1. Thema Markert1: Popularitätsranking für soziale Medien
   - Markert1a: **Learning to Rank** für Popularitätsranking
   - Markert1b: Kohärenz für Popularitätsranking
2. Thema Markert2: Automatische Erkennung von innovativen Metaphern mit **Knowledge Graph Embeddings**

Markert1: Popularitätsranking für soziale Medien

Aufgabe

Gegeben einen Social Media Post, sage dessen Popularität voraus

Variationen:

- Maße: Likes, retweets, #comments, Upvotes
- Wann: Vor Posting (Cold Start), nach einer gewissen Zeit (Netzwerkeffekte)
- Wie: Metadaten, Inhalt, Bilder ...

Metadaten vs NLP

Can't sleep going through old pics. Posted some on my ig stories

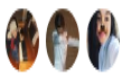
1:30 AM - 29 Mar 2019

314 Retweets **10,198** Likes

Reminiscing on old photos at 5 am because you can't go back to sleep 💤

4:47 AM - 11 Apr 2019

3 Likes



3

Metadaten vs NLP

Bei sozialen Netzwerken mit starker Followerstruktur sind Metadaten mit NLP-Methoden schwer zu schlagen.

Gegensatz: Frage-Antwort-Foren, Narrative Foren

Aus Reddit's Explain Like I'm 5: *Why heart muscles can work 24/7 nonstop but other muscles like biceps can't and get tired to they point the can't contract more?*

- A1: *Skeletal muscles get tired and lose contractility because you deplete the muscles' stores of glycogen (basically the stored form of sugar). They lose their easy to access energy source and stop working as well. Cardiac muscle feeds primarily on circulating fats, and are designed to harness this very efficiently. Because they aren't relying on stored energy but instead circulating energy, they don't need to worry about depleting their stores.*

Gegensatz: Frage-Antwort-Foren, Narrative Foren

Aus Reddit's Explain Like I'm 5: *Why heart muscles can work 24/7 nonstop but other muscles like biceps can't and get tired to they point the can't contract more?*

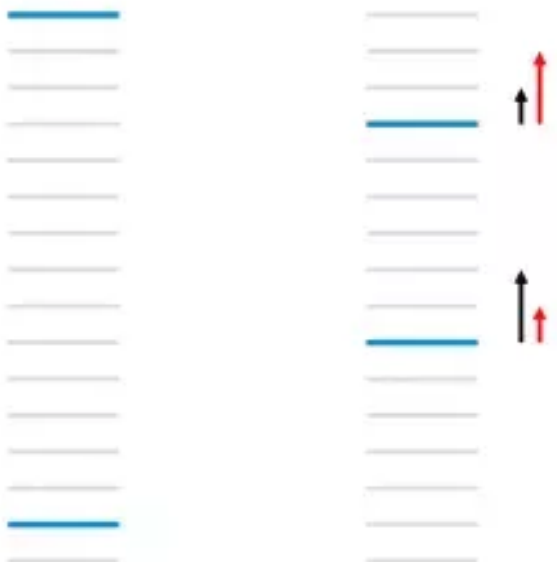
- A2: *Efficiency, to have muscle work indefinitely requires more energy, which requires more food. A human who's every muscle can work flat out indefinitely wouldn't be able to gather enough energy to survive during a food shortage. Also, the heart has a single mostly unchanging task, it too gets tired and fails when overworked.*

Markert1a: Learning to Rank für Popularitätsranking

Bisherige Arbeiten:

- Meist: Optimiere Vorhersage eines einzelnen Posts (Klassifikation, regression)
- Manchmal: Entscheide, welcher Post von einem Paar der bessere ist.

Learning to Rank: Loss function ist paarweise oder listenweise loss function. Direkte Optimierung von Ranking Metriken.



Algorithmen und Toolkits

- RankNet (2005), LambdaRank (2006): Minimiere paarweise Inversionen mit SGD
- TF-Ranking (2018): Scalable Tensor-Flow Library for Learning-to-rank. Implementiert auch groupwise scoring functions.

Hauptziel: Verstehe und experimentiere mit Learning to Rank

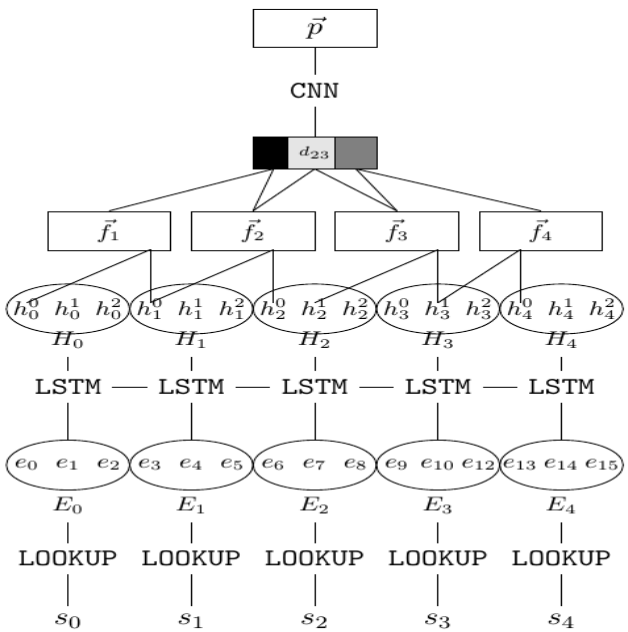
Literatur

- Tan, Lee and Pang (2014). The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. In Proceedings of ACL.
- Hessel, Lee and Mimno (2017). Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In Proceedings of the 26th International Conference on World Wide Web, pp. 927–936.
- TF-Ranking (2018): `https://ai.googleblog.com/2018/12/tf-ranking-scalable-tensorflow-library.html`
- Burges, C. J. (2010). From ranknet to lambdarank to lambdamart: An overview. *Learning, Volume 11*.

Markert1b: Kohärenz und Popularität

Anstatt auf den Algorithmus kann man sich auch auf linguistische Modellierung konzentrieren.

- Für Frage-Antwort-Foren für Laien: Inwieweit spielt **Kohärenz** der Antwort eine Rolle für Popularität?
- Es gibt neuronale und graphenbasierte Kohärenzmodelle. Beispiel: Mesgar und Strube (NAACL, 2018): A neural local coherence model for text quality assessment.



Markert2: Erkennung von innovativen Metaphern

> *"**Trope**: [...] jede Form der Rede, die das Gemeinte nicht direkt und sachlich durch das eigentl. Wort ausspricht, sondern [...] durch e. Anderes, Naheliegendes, e. ""ubertragenen" Ausdruck wiedergibt."*
>
> *Gero von Wilpert (1989): Sachwörterbuch der Literatur*

Häufig (jeder dritte Satz). Wichtig für Sentiment Mining und Text Simplification.

Innovative vs. Konventionelle Metaphern

- Innovativ: *Because the words started as* **coat-hanger** *to hang pictures on*
- Konventionell: *we all live on* **tight** *budgets*
- Konventionell: *I felt* **down** *yesterday*

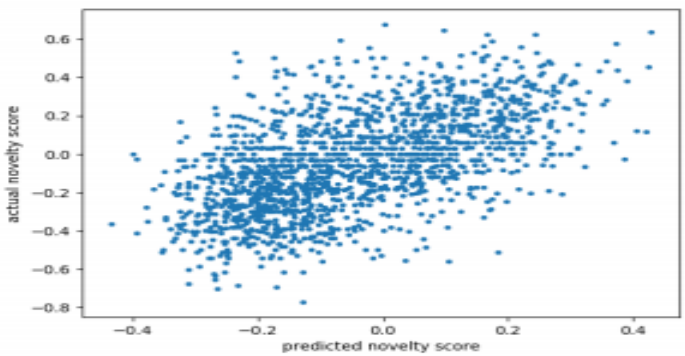
Die meisten bisherigen Arbeiten zur Metaphernerkennung unterscheiden nicht zwischen den beiden Metapherntypen.

Dinh et al EMNLP 2018

Korpora

- VU Amsterdam Metaphor Corpus: Jedes Wort nach Metaphern annotiert.
- Dinh et al erweitern VU Corpus mit Novelty Score.
- +0.75 *Because the words started as* **coat-hanger** *to hang pictures on* (on children's books)

Performanz BISLTM. Nur für Verben.



Trainingsdaten (viel) zu klein → Wissen von außen oder unüberwachte Methoden

Projektidee(n): Metaphor Novelty Scoring mit Knowledge graph embeddings

ConceptNet:

Ausgedrückt in Tripeln: [head,rel,tail] wie [person capable-of eat].

Sollte helfen bei Metaphern wie *This house eats all my energy.*

Projektidee: Knowledge Graph Embeddings

- KGE: Gegeben eine Menge von Tripeln [h, r,t] werden niedrigdimensional embeddings der Entitäten und Relationen gelernt. Zum Beispiel mit Constraints wie $h + r = t$ (Bordes et al 2013)).
- Die entsprechenden embeddings können dann zur Prädiktion von "Metaphor Novelty Scores" verwendet werden.

Ressourcen und Literatur

- Ressourcen
  - VU Amsterdam Metaphor Corpus: `http://ota.ahds.ac.uk/headers/2541.xml`
  - Dinh et al Korpuserweiterung: `https://github.com/UKPLab/emnlp2018-novel-metaphors`
- Papiere:
  - Dinh et al (2018). Weeding out conventionalized metaphors: A Corpus of Novel Metaphor Annotations. In Proceedings of EMNLP 2018
  - Mao et al (2018). Word embedding and WordNet Based Metaphor Identification and Interpretation. In Proceedings of ACL 2018.
  - Bordes et al (2013): Translating Embeddings for Modeling Multi-Relational Data. In Proceedings of NIPS 2013.