Wortdisambiguierung (WSD) für Twitter

- Ein Graphenbasierter Ansatz -

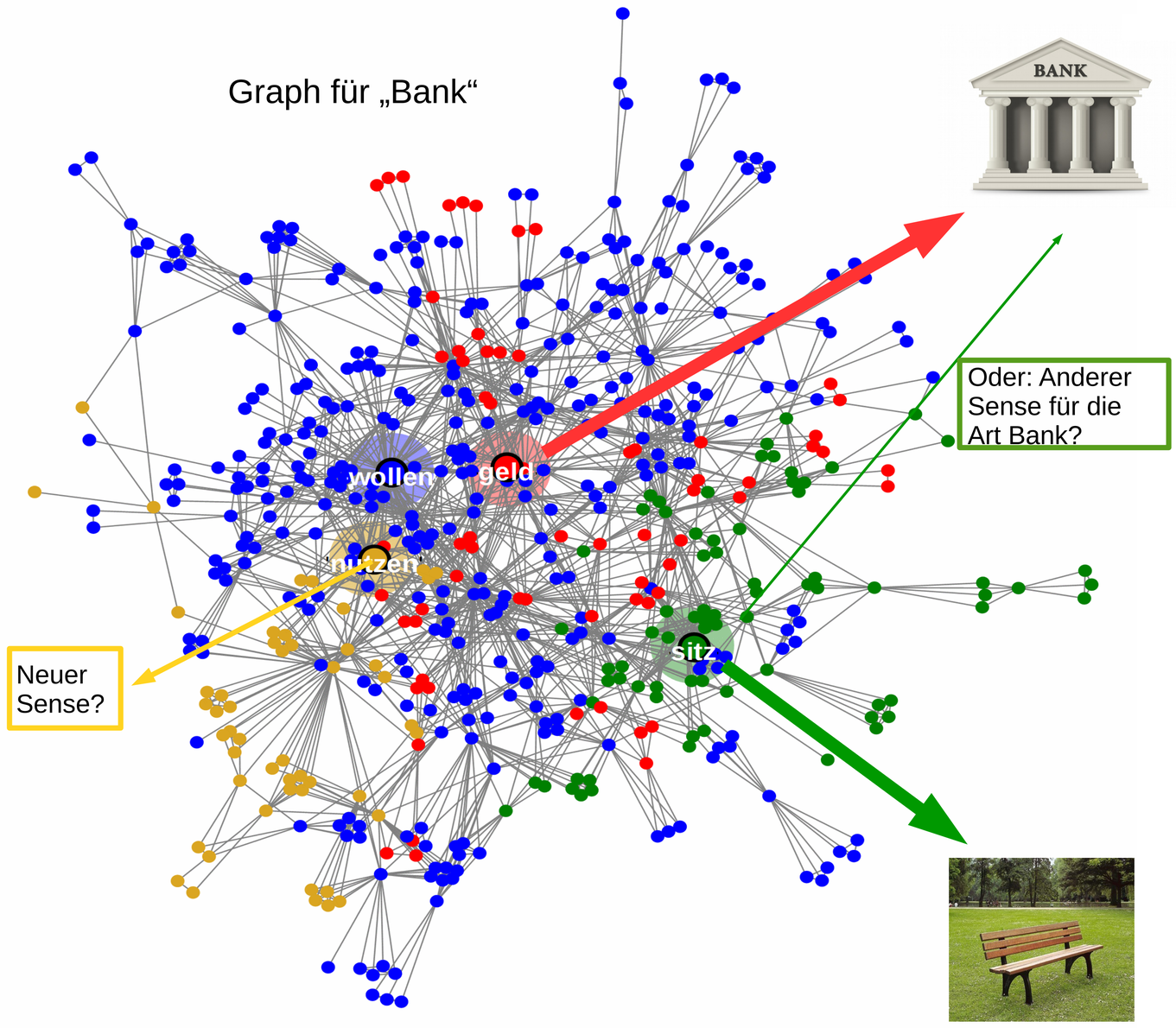
von Sebastian Burst, Arthur Neidlein, Juri Alexander Opitz -

Twitter: **Senses im Wandel, neue Senses entstehen**

Germanet (v 9.0) zu Maus:
1. Kleines graues oder braunes Säugetier mit langem Schwanz, rattenähnlich
2. Computereingabegerät mit Zeigefunktion

Twitter:

Linguistisch motivierte Sensesuche.
Wir suchten in Twitter, Germanet und Duden nach Senses und definierten für Twitter dominierende Senses selbst. Hier: Zusätzlicher Sense für Maus: „Konsename“.



Generelles Problem bei Twitter: **Mangelnder Kontext,** ein Tweet ist auf 140 Zeichen begrenzt.

Wir erweitern den Kontext der Tweets durch Tweets mit gleichem Hashtag, die am selben Tag getweetet wurden.

Wir ermitteln dichte Zentren der Graphen, einmal durch Pagerank und einmal durch die Vorkommenshäufigkeit der Wörter.
Für jeden Knoten wird die kleinste Pfadlänge zu den Zentren berechnet.

Aus einem großen Korpus von Tweets (ca. 1,4 Mio.) wird für jedes ambige Wort ein Graph erstellt.
Knoten sind die Inhaltswörter der Tweets (ohne das Wort selbst) mit dem ambigen Wort.
Kanten werden nach Kookkurrenz der Nachbarknoten gewichtet.

Anwendung auf zu disambiguierende Tweets:
- Für ein zu disambiguierendes Wort werden die Distanzen der Umgebungsinhaltswörter zu den Zentren addiert. Gesamtdistanz = Semantische Nähe zu den Zentren
- Mithilfe eines Goldstandardteils (von Studenten annotierte Tweets) ordnen wir einmalig den Zentren die Wortbedeutungen zu und erstellen eine Mapping-Matrix.
- Mappingmatrix * Nähe zu Zentren = Nähe zu Wortbedeutungen.

Wort	Vork. Gesamt	Senses	Präz MFS	Präz. Sys
Bahn	2750	2	0,56	0,6
Schloss	1600	2	0,99	0,99
Grund	1352	2	0,94	0,94
Eis	790	2	0,78	0,81
Band	700	4	0,83	0,77
Bank	655	2	0,71	0,72
Glas	495	3	0,75	0,76
Erde	475	3	0,59	0,61
Maus	336	3	0,47	0,38
Schlange	310	2	0,87	0,84
Ente	250	2	0,91	0,82
Gras	245	2	0,67	0,53
Leiter	233	2	0,81	0,91
Kippe	221	2	0,96	0,96
Krebs	210	2	0,95	0,91
Schuppen	205	2	0,58	0,6
Platte	203	3	0,39	0,22
Operation	200	2	0,87	0,89
Kiefer	198	2	0,93	1
Stamm	178	3	0,57	0,56
Zelle	150	2	0,64	0,43
Bruch	150	3	0,5	0,75
Ungew. Mittel			0,741	0,727

Ergebnisse:
- Bei Hinzunahme des Hashtagkontexts ergab sich bei einigen Wörtern ein Präzisionsgewinn.
- Leider konnte kein allgemeiner Zusammenhang zwischen mehr Kontext und höherer Präzision festgestellt werden.
- Die starke Most Frequent Sense Baseline konnte bei 10 von 22 Wörtern geschlagen werden, 9 mal schlug die Baseline unser System.

Weitere Erkenntnisse:
- Graphen besitzen viel Potenzial komplexe und sich verändernde Senses z.B. einer Alltagssprache abzubilden.
- Die Auswahl der Zentren ist parameterabhängig. Wir benutzten lediglich eine Mindestdistanz zwischen zwei Zentren, weitere Parameter ließen sich aber implementieren.
- Diese Distanz beeinflusst die Ergebnisse stark, die optimale Distanz variierte von Graph zu Graph.
- Wir wählten im Vorfeld einen Wert aus, durch Tunen dieses Wertes (z. B. mithilfe eines Developmentsets) ließe sich das System evt. stark optimieren!

Kontakt: <nachname>@cl.uni-heidelberg.de