Extending and Exploiting the Entity Graph for Analysis, Classification and Visualization of German Texts

Julia Suter and Michael Strube, Heidelberg Institute for Theoretical Studies

Heidelberg Institute for Theoretical Studies
HITS

In a Nutshell

- Enhancing the entity graph
- Visualizing and analysing German texts using the entity graph
- Text classification with entity graph metrics

Entity Grid — Barzilay and Lapata (2008)

s_1 A little mouse sits in *his* hole. s_2 Although he has spotted some cheese just outside the hole, he doesn't dare to go get it. s_3 What if the *neighbor's* cat is around? s_4 *His* body shivers at the thought of the furry *predator*. s_5 Then, the mouse hears a barking sound close by. s_6 It could only come from the huge dog that sometimes chases the cat around. s_7 Confidently, the mouse crawls out of the hole - and *is promptly caught and eaten by the cat*. s_8 "It's good to be bilingual", purrs the cat.

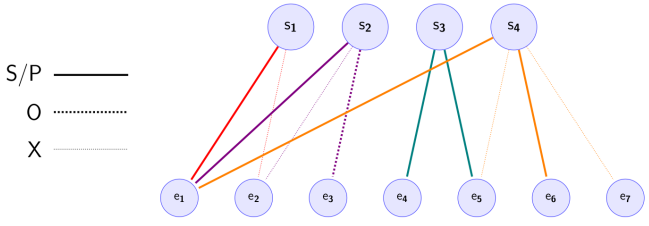
Adjustments to syntactic role categories:

- ✓ Category for possession modifiers
- ✓ Weight reduction for embedded entities

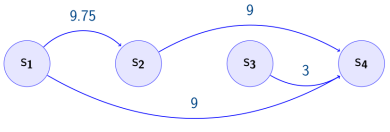
Weights
S/P = 3
O = 2
X = 1

	MOUSE	HOLE	CHEESE	NEIGHBOR	CAT	BODY	THOUGHT	SOUND	DOG
s_1	S	X	–	–	–	–	–	–	–
s_2	S	X/	O	–	–	–	–	–	–
s_3	–	–	–	P	S	–	–	–	–
s_4	P	–	–	–	X	S	X	–	–
s_5	S	–	–	–	–	–	–	O	–
s_6	–	–	–	–	O/	–	–	S	S/
s_7	S	X	–	–	S	–	–	–	–
s_8	–	–	–	–	S	–	–	–	–

Entity Graph — Guinaudeau und Strube (2013)



One-mode projections (P_U and P_W) capture the relations between sentences:



Average out-degree: 5.125

Average out-degree is suitable measure for local coherence

Improvements with Adjusted Syntactic Categories

Evaluation by sentence reordering task on TüBa-D/Z news corpus

Original Entity Graph

	Disc. acc.	Pos. ins.	Ins. acc.
P_U	0.881	0.124	0.102
P_W	0.880	0.142	0.114

Adjusted Entity Graph

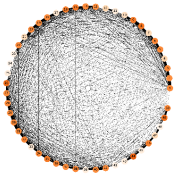
	Disc. acc.	Pos. ins.	Ins. acc.
P_U	0.889*	0.132*	0.105*
P_W	**0.900***	**0.156***	**0.119***

Text Analysis by Graph Metrics

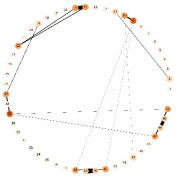
- Visualization of entity graphs reveals differences in entity distribution among different authors and text types
- Graph metrics as features for analysis and classification

- Text samples of 50 sentences extracted from Project Gutenberg
- Syntactic parsing using *ParZu*, coreference resolution using *CorZu*

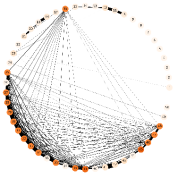
Entity Graph Examples for German Literary Texts



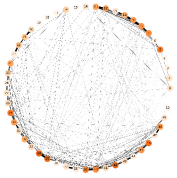
Brothers Grimm
Hans im Glück
Sents. 1-50



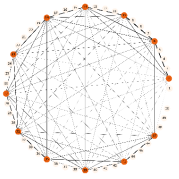
May
Am Rio de la Plata
Sents. 801-850



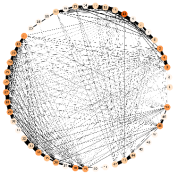
Schnitzler
Die Frau des Weisen
Sents. 151-200



Jean Paul
Jean Pauls Briefsammlung
Sents. 501-550



Tieck
Hexensabatt
Sents. 201-250



Kafka
Das Urteil
Sents. 151-200

Author and Genre Classification

- Support Vector Classifiers
- 24 features extracted from entity graph: *centrality measures, edge weights, diameter, maximum flow, edge betweenness, clustering coefficient*, etc.
- Comparison to 31 syntactic features

Does adding entity graph features improve the syntactic feature system?

Author Classification

	Acc.	F
EG P_W	0.470	0.298
Syntactic	0.740	0.731
+ EG P_W	**0.787***	**0.772***

Genre Classification

	Acc.	F
EG P_W	0.416	0.298
Syntactic	0.572	0.541
+ EG P_W	**0.615**	**0.560**

→ Entity graph features capture information not encoded in syntactic ones

Conclusion

- Possession modifiers category and reduced weights improve entity graph
- Entity graph visualization and graph metrics are suitable for analyzing texts
- Contribution of new information to author classification

Acknowledgements

This work has been funded by the Klaus Tschira Foundation, Heidelberg.