# Information Extraction Topics

CIS, LMU Munich

Winter Semester 2020-2021

Denis Peskov, UMD and CIS

# Organizing Wikipedia

- Wikipedia is a collective data source
  - Different users contribute to the same page
  - But pages should be similar across pages
  - And even across pages
  - Given that Wikipedia and WikiData are both publicly available data sources, can one automatically extract the slots needed for WikiData from Wikipedia?

- Recommended Papers:
- Vrandečić, D. and Krötzsch, M., 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), pp.78-85.
- Färber, M., Bartscherer, F., Menne, C. and Rettinger, A., 2018. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web*, *9*(1), pp.77-129.

# Annotating Dialogue Acts

- Automatically identifying slots and intents within dialogues
  - Dialogues occur naturally in internet data in sources such as forums
  - Can we create systematic annotations for dialogues?

- Recommended Papers:

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C.V. and Meteer, M., 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, *26*(3), pp.339-373.

Budzianowski, P., Wen, T.H., Tseng, B.H., Casanueva, I., Ultes, S., Ramadan, O. and Gašić, M., 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *EMNLP*

# Extracting Structured Document Information

- Much of the data on the internet is created according to a skeleton
  - PDF documents have clear headers
  - Legal documents contain various topics
  - Can we automatically detect what content belongs to which section?

- Recommended Papers:
- Koshorek, O., Cohen, A., Mor, N., Rotman, M. and Berant, J., 2018. Text segmentation as a supervised learning task. *NAACL*
- Barrow, J., Jain, R., Morariu, V., Manjunatha, V., Oard, D.W. and Resnik, P., 2020, July. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 313-322).