Neural De-Duplication and Record Linkage

Expert: Grant Galloway (Amazon Research)
Session: Session 2 (5 July – 6 August)

Record Linkage or Entity Resolution is the task of grouping similar entities across one or more data sources. It is commonly used for improving data quality and integrity, to allow re-use of existing data sources for new studies, and to reduce costs and efforts in data acquisition [1][2]. Applications include matching bibliographic and e-commerce data entities [3], matching business records [4] and predicting duplicate ads [5]. Within Amazon, Record Linkage methods are used to find relationships between products. These relationships are used to drive search and discoverability of products, and to improve the shopping experience on the website.

A key component in many Record Linkage systems is a matching component that determines whether pairs of records refer to the same entity. State of the art systems use Machine Learning models to perform this task. In this project you will explore how recent advances in NLP and Deep Learning improve the the matching problem. You might answer one (or more) of the following questions:

1. Can we learn a model from raw input data (e.g. text or images) to outperform hand-crafted features?
2. How do different word segmentation techniques (such as Byte Pair Encoding [6]) affect the performance?
3. Can we learn unsupervised representations and apply transfer learning for this domain? Does the pre-trained model and active learning drastically reduce the amount of annotated data we need? [7]
4. How do deep learning based approaches compare with other machine learning methods such as Random Forest, Decision Tree, SVM, etc.?

Small, manageable datasets can be found at:
https://hpi.de/naumann/projects/repeatability/datasets.html
These datasets contain mostly textual data only.

Useful Courses: Python Programming, Machine Learning in Python

References:
[1] Christen, Peter. "A survey of indexing techniques for scalable record linkage and deduplication." IEEE transactions on knowledge and data engineering 24.9 (2012): 1537-1555.
[2] Elmagarmid, Ahmed K., Panagiotis G. Ipeirotis, and Vassilios S. Verykios. "Duplicate record detection: A survey." IEEE Transactions on knowledge and data engineering 19.1 (2007): 1-16.
[3] Köpcke, Hanna, Andreas Thor, and Erhard Rahm. "Evaluation of entity resolution approaches on real-world match problems." Proceedings of the VLDB Endowment 3.1-2 (2010): 484-493.
[4] Gruenheid, Anja, Xin Luna Dong, and Divesh Srivastava. "Incremental record linkage." Proceedings of the VLDB Endowment 7.9 (2014): 697-708.
[5] https://www.kaggle.com/c/avito-duplicate-ads-detection/data
[6] Rico Sennrich, Barry Haddow, Alexandra Birch, "Neural Machine Translation of Rare Words with Subword Units", Volume: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1715–1725
[7] Jungo Kasai et al. "Low-resource Deep Entity Resolution with Transfer and Active Learning", Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5851–5861