Fall 2018 Summary

Ahmed El-Kishky
Parsimonious Morpheme Segmentation with an Application to Embedding

• Submitted a paper to SDM for an unsupervised subword mining algorithm:
  1. Unsupervised, so works on many concatenative languages (English, German, Turkish)
  2. Hierarchical so creates subwords at multiple granularities
  3. Method recovers ground-truth morphemes beyond previous State of the Art
  4. Improves downstream word embeddings on:
     1. Word Similarity
     2. Word Analogies
     3. Language Modeling

• Ahmed El-Kishky, Frank Xu, Aston Zhang, Jiawei Han, *Parsimonious Morpheme Segmentation with an Application to Word Embeddings*. Under Review (SDM’19)
Constrained Sequence-to-Sequence Classification for Semitic Root Extraction

• Submitting a paper to NAACL for supervised Semitic root extraction:
  1. Identifies non-concatenative roots (subwords that can skip characters / not contiguous)
  2. Extracts roots better than state-of-the-art on Arabic and Hebrew.
  3. Introduce modification of sequence-to-sequence model for multi-class classification.
     • Works really well when input and label share the same semantic space / label space is sequential.
  4. Improves downstream word embeddings on:
     1. Word Similarity
     2. Word Analogies
     3. Language Modeling

• Ahmed El-Kishky, Xingyu Fu, Aseel Adawood, Nahil Sobh, Clare Voss and Jiawei Han,
  *Constrained Sequence-to-Sequence Classification for Semitic Root Extraction. Submission (NAACL’19)*
Collaborations

• Doris Xin, Ahmed El-Kishky, De Liao, Brandon Norick, and Jiawei Han, "Active Learning on Heterogeneous Information Networks: A Multi-armed Bandit Approach", in Proc of 2018 IEEE Int. Conf. on Data Mining (ICDM'18), Singapore, Nov. 2018

• Qi Zhu, Xiang Ren, Jingbo Shang, Yu Zhang, Ahmed El-Kishky and Jiawei Han, "Integrating Local Context and Global Cohesiveness for Open Information Extraction", in Proc. 2019 ACM Int. Conf. on Web Search and Data Mining (WSDM'19), Melbourne, Australia, Feb. 2019
Zero-Shot Ultra-Fine Grained Entity Typing

• Work in progress: Aiming for KDD submission
  1. Many types (PubMed=8K types, Wikipedia+FreeBase=1.5K types)
  2. Using subword-embeddings improves typing (especially PubMed)
  3. Putting input and labels in same embedding space addresses large number of labels.

• Message me if you would like to discuss / collaborate 😊.