Identifying Emerging Technology: A Neural Network Based Solution

Jin Mao, Chao Ma, Zhentao Liang

Center for Studies of Information Resources, Wuhan University

Oct. 17, 2019

Problem formulation

- Input:

  Bibliographic records of a domain

  Technical terms and their features in the first 8 years (time-series) according to the records

- Features:
  - Document frequency (DF);
  - Total number of authors (AN), citations (CN), references (RN), and funding (FN) each year;
  - Accumulated number of authors (AA), citations (AC), references (AR), and funding (AF);
  - **Emerging Score** of each year

- Goal:

  predict the **Emerging Score** of **technical terms** in the 10th year.

Methodology

- Technical terms extraction

  The Termolator (Meyers et. al, 2018) was used, 10,000 technical terms were extracted from processed text fields, including title, abstract, keyword, and Keyword Plus.

- Emerging Score

  Consider both **document frequency** and **growth rate**:

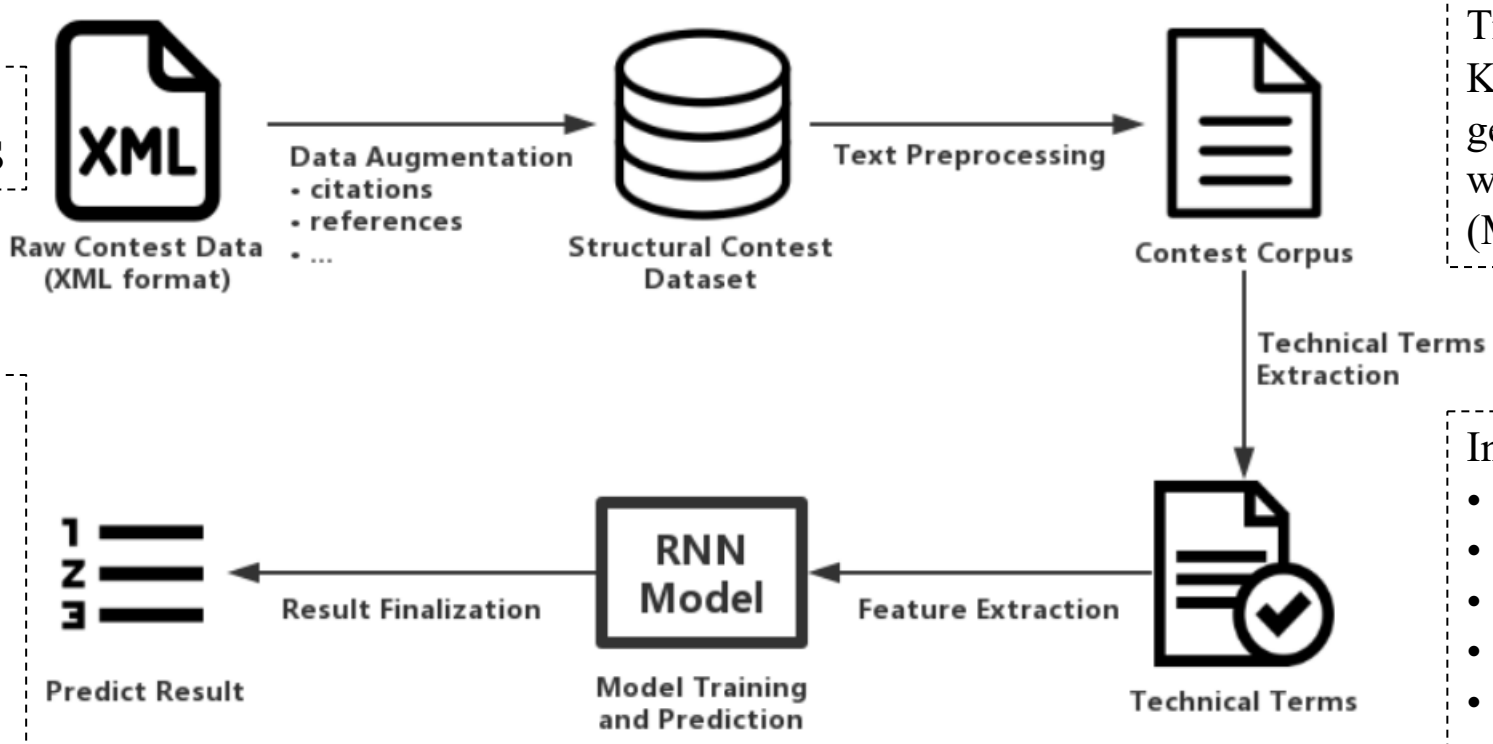
  (1) $df_y^{t_i} = \#(t_i) + \delta$; (2) $\mathrm{gr}_y^{t_i} = \frac{df_y^{t_i} + df_{y-1}^{t_i}}{df_{y-2}^{t_i} + df_{y-3}^{t_i}}$; (3) $\mathrm{es}_y^{t_i} = \log\left(df_{y-2}^{t_i} + df_{y-3}^{t_i}\right) * \mathrm{gr}_y^{t_i}$

- LSTM model
  - Optimized the mean squared error (MSE) between predicted and true Emerging Score
  - Normalized discounted cumulative gain (NDCG) @10 was used for model selection

Meyers, A. L., He, Y., Glass, Z., Ortega, J., Liao, S., Grieve-Smith, A., ... & Babko-Malaya, O. (2018). The Termolator: terminology recognition based on chunking, statistical and search-based scores. Frontiers in Research Metrics and Analytics, 3, 19.

Experiment detail

All bibliographic fields were completed by searching WoS



Title, abstract, keyword and Keyword Plus were used to generate technical terms, with the help of Termolator (Meyers et. al, 2018)

Results:

1. Operon
2. Heterologous gene
3. Cell-free protein
4. Genetic oscillator
5. Mevalonate
6. Metabolic engineering
7. Tumor necrosis
8. Synthetic biology approach
9. Aptamer
10. Mycoplasma

- LSTM cells
- MSE between predicted emerging score and ground truth
- NDGC@10 for model selection (best 0.74)

Input features:

- Document frequency (DF)
- Author count (AC)
- Citation count (CC)
- Reference count (RC)
- Funding count (FC)
- Accumulated DF
- Accumulated AC
- Accumulated CC
- Accumulated RC
- Accumulated FC
- Emerging score

Thanks!