Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search

Zhuyun Dai, Chenyan Xiong, Jamie Callan, Zhiyuan Liu
Carnegie Mellon University, Tsinghua University

• Queries and documents often match at n-gram level
  • Qry: `atypical squamous cells’ Doc: “to prevent cervical cancer...”
• Traditional IR: Exact matching n-grams (Vocabulary mismatch)
• K-NRM (SIGIR17): A neural ranking architecture that learns word-to-word soft-match patterns with kernel pooling

Conv-KNRM: a neural ranker for soft-matching n-grams in ad-hoc search

Convolution
Compose n-gram embeddings from adjacent words’ embeddings

Cross-Matching
Soft match n-grams of different lengths (e.g. Query Trigrams to Document Bigrams)

KernelPooling
Extract multi-level soft-match features (e.g. Exact match, strong match, weak match...)
Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search

Zhuyun Dai, Chenyan Xiong, Jamie Callan
Carnegie Mellon University

Zhiyuan Liu
Tsinghua University

• Effective: 30%+ over strong baselines
• Cross-matching n-grams of different length is the key
  • Qry : “Convolutional Neural Networks” Doc: “Deep Learning”
• Generalizable across search domains

Search Log (End-to-End)
• Learn Word Embedding
• Learn CNN
• Learn LeToR weights

TREC Web Track (Adapt)
• Apply Word Embedding
• Apply CNN
• Learn LeToR weights

The n-gram soft-matching patterns are effective in related domains.