Distributed Non-Parametric Representations for Vital Filtering

UW @ TREC KBA

Ignacio Cano, Sameer Singh, Carlos Guestrin

Problem
- Exponential increase of information
- Streaming corpora of text docs
- Critical to detect relevant events
- Incorporate info to entities timely

Document Categories
- Referent:
  • VITAL: “Barack Obama has been elected President”
  • NON-VITAL: “Barack Obama was born on August 4th, 1961”
- Unknown: “Barack is a great father and a better husband”
- Non-Referent: “Barack Ferrazzano provides a wide range of business-oriented legal”

Method

1. Word Embedding Representations of Documents
  - Dense, low dimensional vectors representations of words
  - Address sparsity in BOW models
  - Efficient training on massive corpora
  - Encode syntactic and semantic properties of words
  - Represent documents with mean embedding vector

2. Multiple Embeddings for Entity Contexts (Clustering)
  - Represent an entity with an embedding that captures what we’ve seen about that entity.
  - Distance of a new document to the embedding is a good indicator of novelty of the document.
  - Entities are mentioned in multiple contexts, having a single embedding may conflate the topics.
  - Better to have multiple embeddings:
    • Advantages of using embeddings
    • Still have a precise context representation
  - Represent clusters with mean word embedding vector of all documents assigned to that cluster.
  - Assume each document belongs to single cluster.
  - Do not need to know number of clusters beforehand.

3. Staleness Measure
  - Document is vital when it provides new, timely information to an entity profile.
  - Current representation cannot capture timeliness.
  - Documents close to existent clusters may contain novel information.
  - Staleness intends to capture the timeliness of information (temporal dynamics)

Results
- Preprocess corpus using exact string matches to target entity names.
- Pre-trained embedding vectors on part of the Google News dataset \(V=3M, d=300\)

Accelerate & Create
- Browser-based visualization prototype with interactive time-series controls.
- Document view shows the distribution of vital vs. non-vital documents over time.
- Topic view shows the evolution of topic clusters for a particular entity.
- User can select time ranges to explore over.
- Understand topics using lists of similar words.

Future Work
- Experiments with more datasets
- Learn alpha
- Explore streaming clustering algorithms
- Study more alternate visualizations