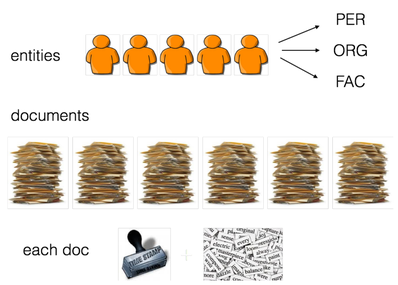
Distributed Non-Parametric Representations for Vital Filtering
UW @ TREC KBA

Ignacio Cano, Sameer Singh, Carlos Guestrin

University of Washington Computer Science & Engineering

Problem

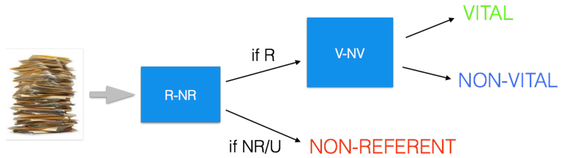
- Exponential increase of information
- Streaming corpora of text docs
- Critical to detect relevant events
- Incorporate info to entities timely



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

france

Word	Cosine distance
spain	0.637530
french	0.632606
germany	0.631435
europe	0.626425
italy	0.625796
england	0.612078
european	0.607490

paris

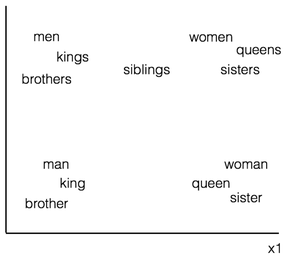
Word	Cosine distance
heidi	0.559217
london	0.555578
france	0.550079
dubai	0.553233
samuel	0.549419
hilton	0.547287
rome	0.546584
toronto	0.545715
las_vegas	0.544174

germany

Word	Cosine distance
german	0.680957
europe	0.678122
european	0.650211
sweden	0.638424
switzerland	0.636213
austria	0.632552
france	0.631436

is

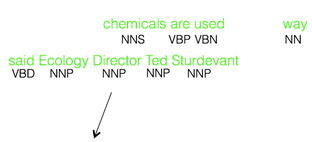
Word	Cosine distance
was	0.654974
isn't a	0.643952
seems	0.634030
Is	0.608597
becomes	0.584194
appears	0.582290
remains	0.579694



- Represent documents with mean embedding vector

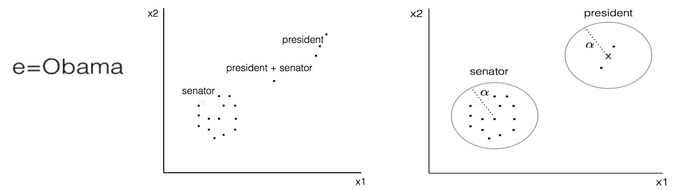
e = Ted Sturdevant

d = ... So, it comes as good news that state officials are making significant progress in determining which chemicals pollute Puget Sound and in identifying where they come from. A report last week from the Department of Ecology casts a wide net over culprits. Most toxic chemicals are used in some way by all of us, said Ecology Director Ted Sturdevant. They are in our homes and gardens...



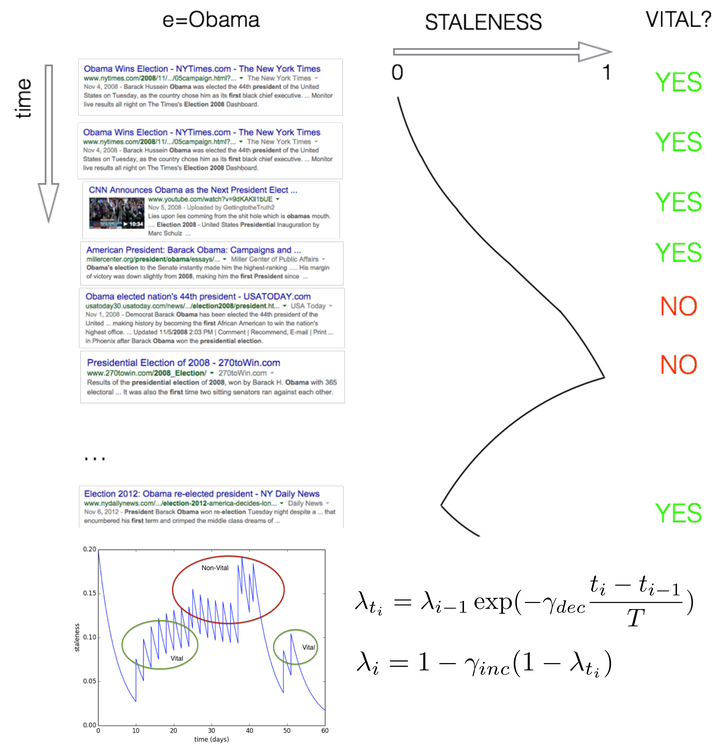
$$\vec{V}_d = 1/7\ (\vec{V}_{chemicals} + \vec{V}_{are} + \vec{V}_{used} + \vec{V}_{way} + \vec{V}_{said} + \vec{V}_{ecology} + \vec{V}_{director})$$

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



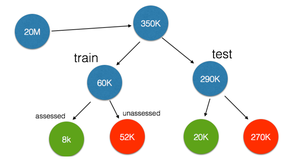
$$\lambda_{t_i} = \lambda_{i-1} \exp(-\gamma_{dec} \frac{t_i - t_{i-1}}{T})$$

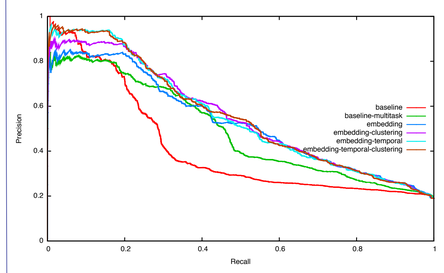
$$\lambda_i = 1 - \gamma_{inc}(1 - \lambda_{t_i})$$

- 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

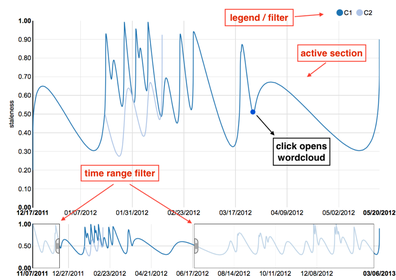


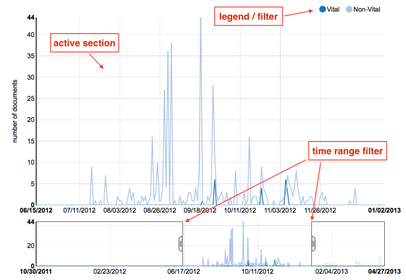


Model	Vital Only	
	Micro F1	Macro F1
Baseline	0.355	0.317
Baseline Multi-task	0.492	0.385
Embedding	0.534	0.409
Embedding-Temporal	0.519	0.397
Embedding-Clustering	0.523	0.403
Embedding-Temporal-Clustering	**0.538**	**0.412**

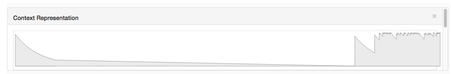
Accelerate & Create

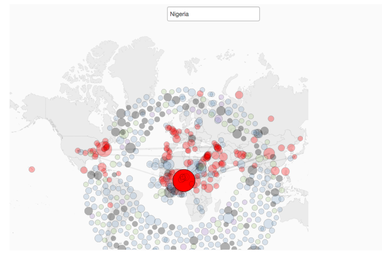
- Browser-based visualization prototype with interactive time-series controls.
- Document view shows the distribution of vital vs. non-vital documents over time.
- Topic view shows the evolution of topic clusters for a particular entity.
- User can select time ranges to explore over.
- Understand topics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





- Explore other alternate visualizations.





Future Work

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