Semantic Matching in Natural Language Text At Scale

Ehsan Kamalloo

January 10, 2019

Querying Beyond Text Surface

- Natural language text is prevalent throughout the Web
- **Challenge:** How to search over enormous size of content rich data
- Traditional IR techniques (e.g., tf-idf [Salton & Buckley, 1975])
    - Ignore rich interactions among words
    - Do not work well in finer granularities than documents

Problems

- **Lexical:** Small word overlap between the query and the target sentences
- **Syntactic:** Different syntactic structures

- Example: *evidence of life in space*
  - *"Alien life looms? Newly discovered exoplanet may be best candidate, experts say."* **Fox**
  - *"Interstellar object may have been alien probe, Harvard paper argues."* **CNN**

Research Goals

- Can we use text meaning rather than surface forms?
    - What schemes have been proposed in the literature to represent meaning?
    - What are the limitations of these schemes?
- Can semantic representation models be incorporated at massive scale?
- Which IR techniques are suitable for scaling up semantic matching?
    - What shortcomings might these techniques have?

Research Goals

- Can we use text meaning rather than surface forms?
  - What schemes have been proposed in the literature to represent meaning?
  - What are the limitations of these schemes?
- Can semantic representation models be incorporated at massive scale?
- Which IR techniques are suitable for scaling up semantic matching?
  - What shortcomings might these techniques have?

Natural Language Semantics

- What humans perceive from text
  - **Example:** *the discovery of liquid water on Mars*
  - *"A large body of water on Mars is detected, raising the potential for alien life."* **Daily Express**
  - *"Scientists have made a huge breakthrough in the search for life on Mars after they discovered what appears to be an existing lake of water."* **NYTimes**
- The argument structure in text
  - who did what, where, when, how, and why

Semantic Representations

- What to represent as meaning [Abend & Rappaport, ACL’16]
  - Events: *Information about occurrence of something*
  - Temporal Relations: *Time-related information*
  - Discourse Relations: *Relations between semantic units*
  - Spatial Relations: *Geographical references*
  - Coreference: *Different mentions of the same entity*
- Representation schemes in this work
  - Shallow Representation Forms
  - Abstract Meaning Representations
  - Embedding Vectors

Shallow Representation Forms

- Define the predicate-argument information using *semantic roles*
- Annotations can be found in FrameNet [Ruppenhofer *et al.*, 2006] and PropBank [Palmer *et al.*, 2005]

- **Example** (culminating clause from the NYTimes news)

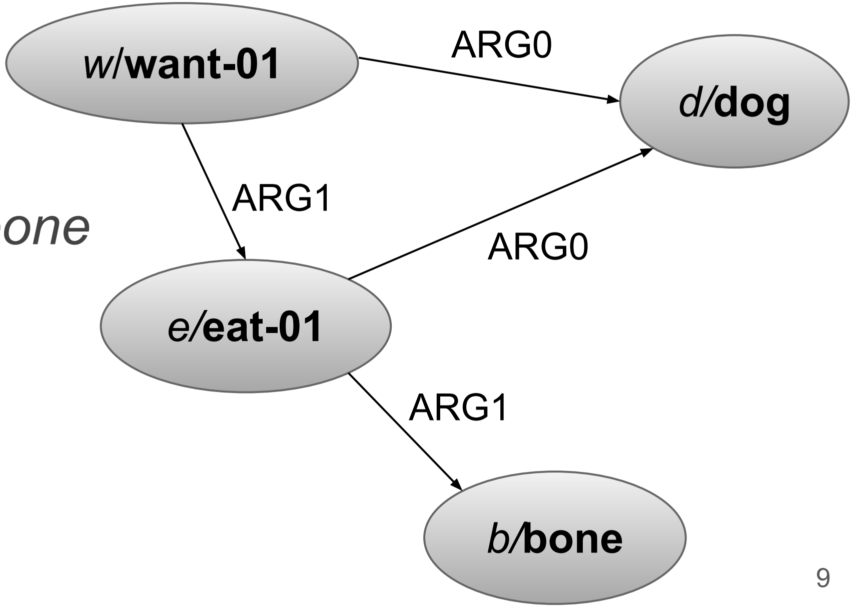
… they **discovered** what appears to be an existing lake of water.

Becoming_Aware

Cognizer **Phenomenon**

Abstract Meaning Representation (AMR)

- Consolidates semantic understanding tasks [Banarescu *et al.*, 2013]
  - Semantic role labeling
  - Named entity recognition
  - Coreference resolution
- Directed Acyclic Graph
- Example: *The dog wants to eat the bone*



Embedding Vectors

- Map semantic information into a latent low-dimensional space
- Inspired from word-level embeddings: Represent each word by a single vector
    - Neural language model [Bengio *et al.*, JMLR’03]: *Slide a window through text to predict the next word*
    - Word2vec [Mikolov *et al.*, NeurIPS’13]: *Given a center word, predict the context words* (Skip-gram model)

Sentence-level Embeddings

- **Unsupervised models:** Based on compositionality of sentences in large text corpora
  - Predict sentences around a sentence (Skip-thought Vectors) [Kiros *et al.*, NeurIPS'15]
- **Supervised models:** Obtained from training on labelled data collected for downstream tasks (i.e., textual entailment in InferSent [Conneau et al., EMNLP'17])
  - **Multi-task learning:** one encoder for many downstream tasks to overcome inductive bias
    - GenSen [Subramanian et al, ICLR'18] and Universal Sentence Encoder [Cer et al., arXiv'18]

Pre-trained Language Models

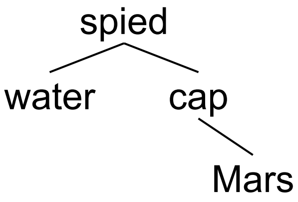
- Learn representations through training a language model
  - **ELMo [Peters et al., NAACL'18]:** Represent a word within its context
  - **OpenAI Fine-tuned Model [Radford et al., arXiv'18]:** Learn a language model using the Transformer model [Vaswani et al., NeurIPS'17]
  - **BERT [Devlin et al., arXiv'18]:** Learn a masked language model (i.e., predict a randomly masked word inside a sentence) using a bidirectional Transformer model

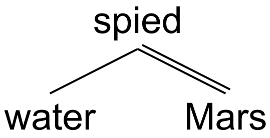
Research Goals

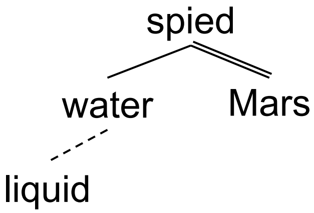
- Can we use text meaning rather than surface forms?
  - What schemes have been proposed in the literature to represent meaning?
  - What are the limitations of these schemes?
- Can semantic representation models be incorporated at massive scale?
- Which IR techniques are suitable for scaling up semantic matching?
  - What shortcomings might these techniques have?

Access Methods over Semantic Representations

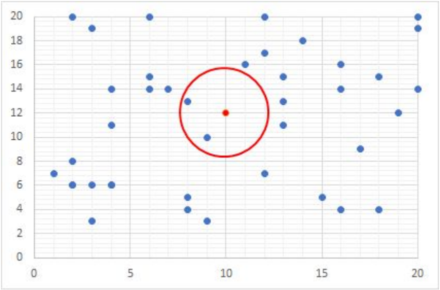
- Annotated Trees (AMRs)
  - Tree Pattern
  - Twig Pattern
  - Generalized Tree Pattern

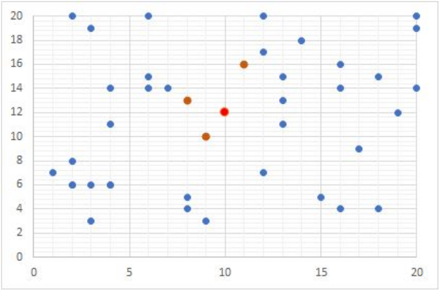






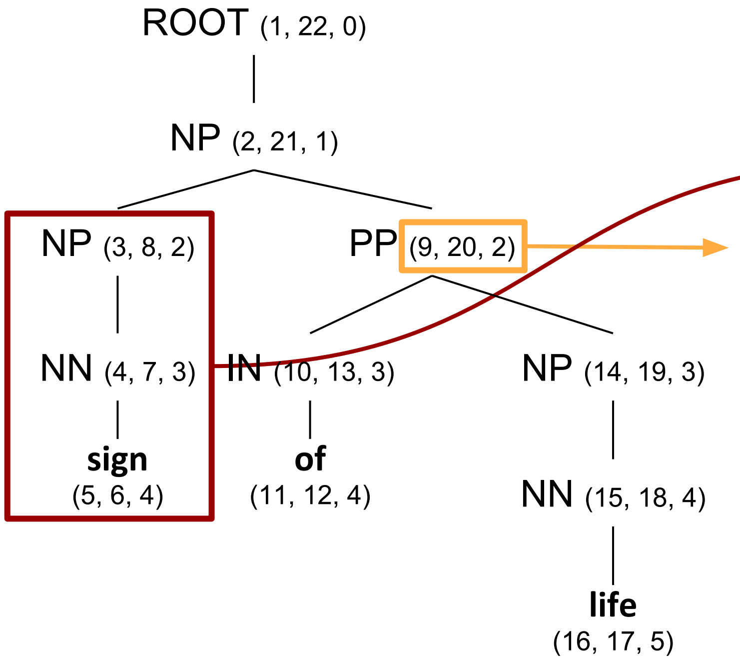
- Multidimensional Data (Embedding Vectors)
  - Range queries
  - *k*-Nearest neighbor queries





Querying over Annotated Trees

- **Subtree Index [Chubak & Rafiei, VLDB'12]:** Interval coding schemes designed for low branching factor trees



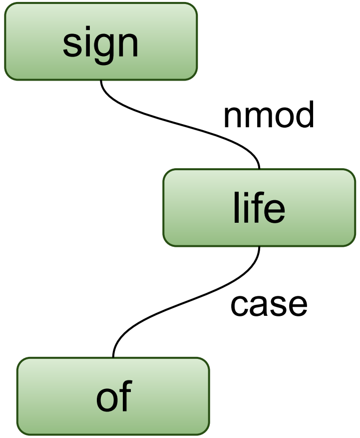
Inverted index (1 .. *mss*)
`NP(NN(sign)) -> <`*`tid`*`, 3,` **`(3, 8, 2), (4, 7, 3), (5, 6, 4)`**`>`

<pre, post, level>

Query: *evidence of life in space*
PP(IN(of)), NP(NN(life)), PP(IN)(NP), ...

Querying over Annotated Trees

- **Subtree Index [Chubak & Rafiei, VLDB'12]:** Interval coding schemes designed for low branching factor trees
- **Koko [Wang *et al.*, VLDB'18]:** Inverted index with interval coding for terms + Hierarchy index to access the tree structure (Closure tables)

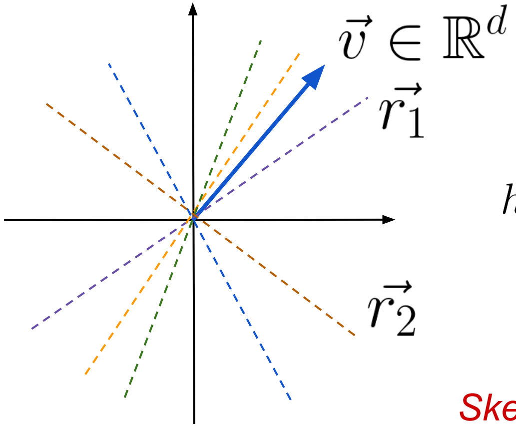


Hierarchy index

```
/root -> sign<sentence_id, token_id, (pre, post, 0)>
/root/nmod -> ...
/root/nmod/case -> ...
```

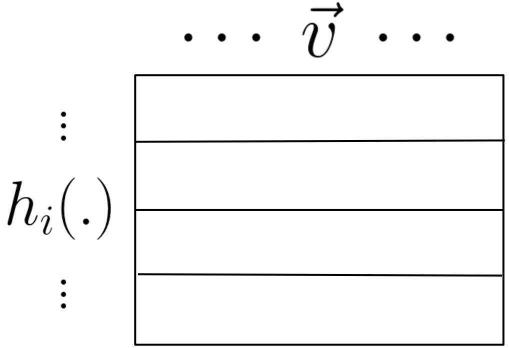
Querying over Multidimensional Data

- Tree structures (k-d-trees and R-trees and their variants)
  - Fall short in high dimensions (curse of dimensionality)
- Locality Sensitive Hashing (LSH)



$$h_i(.) = \begin{cases} +1, & \vec{v} \cdot \vec{r_i} \geq 0 \\ -1, & \vec{v} \cdot \vec{r_i} < 0 \end{cases}$$

Sketches



Generate Candidates

Proposal: Part 1

Enhanced Semantic Representations

- Sentences tend to carry a host of information (meaning inventories)

Coordinating Conjunction

S = "*Watt's original low-pressure designs were able to deliver duty as high as 25 million, but averaged about 17.*" Taken from SQuAD 2.0 [Rajpurkar *et al.*, ACL'18]

S1 = Watt's original low-pressure designs were able to deliver duty as high as 25 million.
S2 = Watt's original low-pressure designs were able to deliver duty averaged 17.

What was the *average* duty of a low-pressure Watt engine? **S2 / S1 / S** (ELMo)
What was the *maximum* duty of a low-pressure Watt engine? **S1 / S / S2** (ELMo)

Proposal: Part 1

Enhanced Semantic Representations

- Sentences tend to carry a host of information (meaning inventories)
- Challenge
    - Find a proper decomposition technique

Proposal: Part 2

Querying over Meaning Inventories

- How to match query representation with inventories of a single sentence
  - Average/Maximum/Minimum over all-pair similarities similar to linkage strategies in hierarchical clustering
- Probing for efficient indexing strategies

Timeline

- Incorporate discourse units to model meaning inventories ~2-3 months
- Probe for effective segmentation strategies ~3 months
- Speed up the retrieval process ~5-6 months
- Build a query language ~1-2 months
- Wrap up and write the thesis ~3 months

Summary

- **Goal:** Improve upon search mechanisms over enormous size of content rich data
- **Problem:** Lexical information and Syntactic structures would not necessarily help
- **Proposed Approach:** Leverage semantic representation of text in searching
- **Anticipated Contributions:**
  - Extract most of the meanings hidden in a sentence
  - Matching strategies for multiple representations