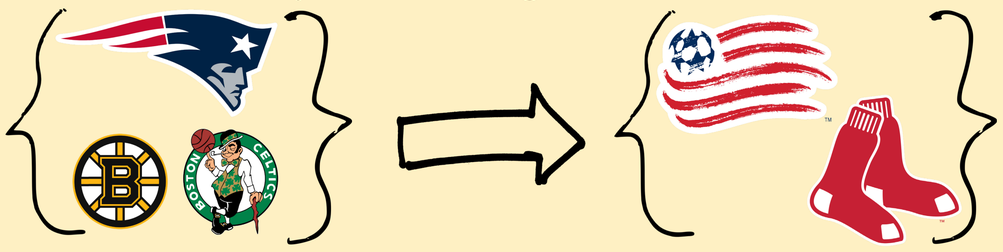
Corpus-based Set Expansion with Lexical Features and Distributed Representations

Puxuan Yu, Zhiqi Huang,
Razieh Rahimi, and James Allan

UMassAmherst | College of Information & Computer Sciences
Center for Intelligent Information Retrieval

Problem and Terminology Statement:

Corpus-based Set Expansion: finding "sibling" entities of user-provided "seed" entities from plain text. Useful for: • QA • Query suggestion • Taxonomy & knowledge base construction ...



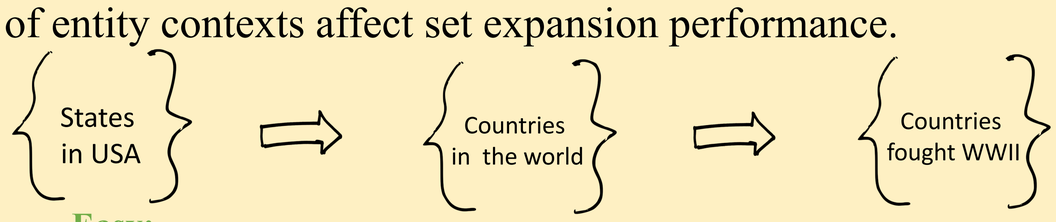
"Major League Sports Teams in New England"

Entity Mention — Lexical Feature (Skip-gram)

"Who scored the first goal in Boston Bruins history?"

Key findings:

1. Sole lexical feature selection produces ranked list of entities with low precision, because the stronger positional constraint is difficult to meet. Our hybrid approach `CaSE` which re-ranks based on semantic similarity between candidate and input entities can significantly improve expansion accuracy.
2. Inclusion relation between entity sets and discrimination power of entity contexts affect set expansion performance.



Easy:
- Frequent entities;
- Distinct contexts

(e.g. "senator from ___")

Normal: Contexts are difficult to distinguish from other sets

Hard: Cannot detect user's intent behind the query

Approach:

Expand seed entities by semantically related entities that frequently share important contexts with seeds.

Input: `ent2ctx, seeds`
Output: `cand_score`

```
cand_score = {}
ctx_union = U(ent2ctx(seed) for seed in seeds)
for ctx in ctx_union:
  for seed in seeds:
    w1 = ent_weight_in_ctx(seed, ctx) # ctx relatedness
    for cand, cand_freq in topn_ents_in_ctx(ctx):
      w2 = cos_sim(seed, cand) # entity semantic similarity
      cand_score[cand] += w1 * w2 * smooth(cand_freq)
return cand_score
```

Tricks:
- Concave "smoothing" function over co-occurrence count
- Increasing and strictly positive function over cos_sim()

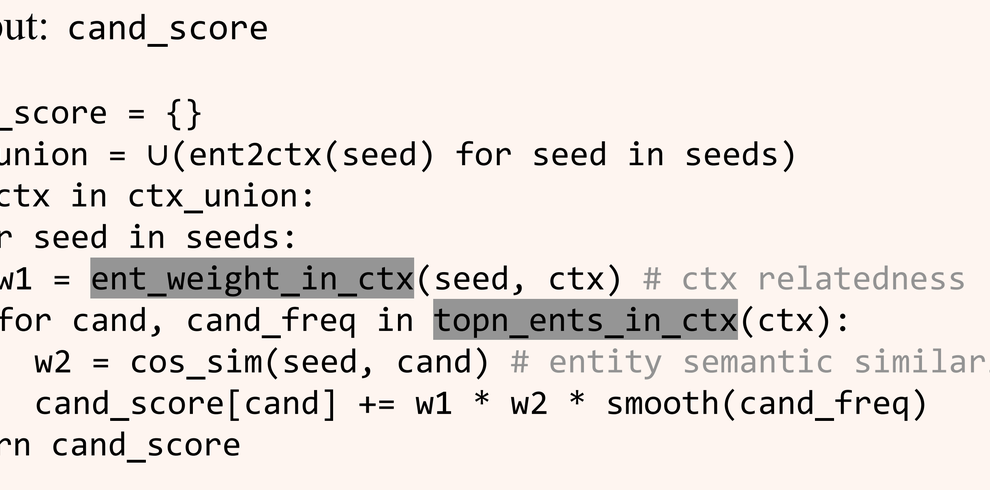
Analysis:

Q: Why hybrid approach outperforms individual methods (context selection alone or embedding KNN alone):
A:
- Skip-gram feature selection poses strong positional constraint but may hurt infrequent entity sets.
- Embedding approaches tend to find topically related but non-substitutable words and phrases.
- Two perspectives complement one another.

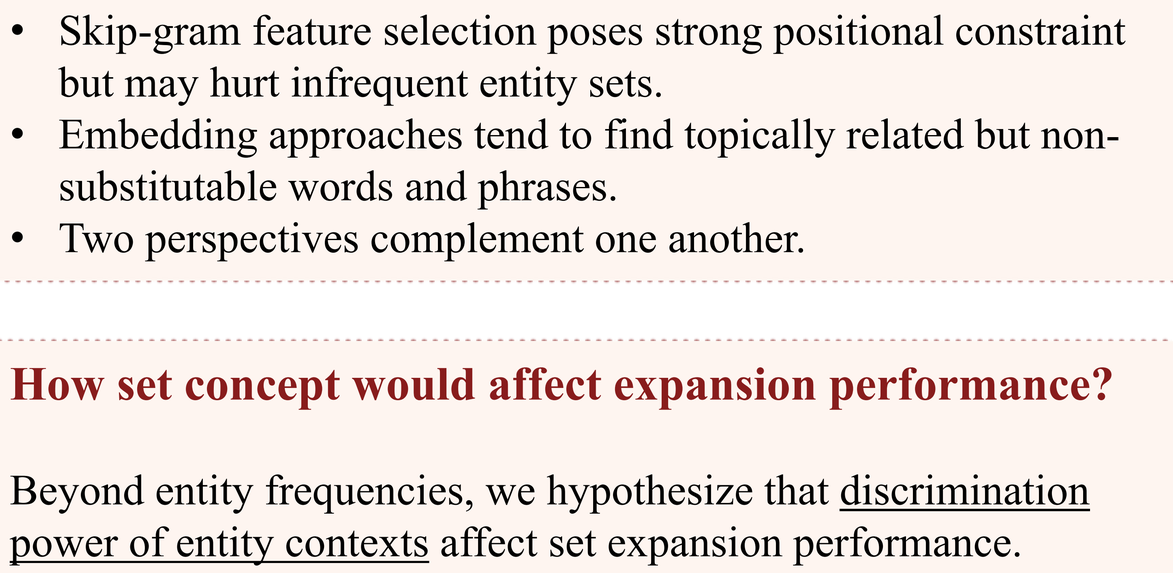
How set concept would affect expansion performance?

Beyond entity frequencies, we hypothesize that discrimination power of entity contexts affect set expansion performance. Therefore, we define k_i as the average number of entities that occur in each context associated with entity e_i, which characterizes *generality* of e_i's contexts. K is the average of k_i normalized over entity frequencies.

Experiments:



- Queries: variable-length entities sampled from 60+ entity sets
- Metric: MAP for queries with the same length
- Methods:
  - Semantic similarity: `Word2Vec, BERT, SetExpander`
  - Lexical feature selection: `SetExpan`
  - Hybrid: `CaSE`



In this manner, a set's K and MAP should be inversely related, but we observed outliners on the bottom left corner, which, by our analysis, turn out to be conceptual subsets of larger entity sets. We believe performance of such sets can be improved by retrieving detailed relevant documents first, then expanding sets from limited text data, which we leave for future work.