Semantic clustering of questions

Maria-Cătălina Mocanu
catalina.mocanu@gmail.com

Problem statement

- Information retrieval:

  one question vs. large amount of data where the answer may be hidden
- Semantic clustering of questions:

  large number of questions vs one persons limited ability to give answers

- Restrictions:

  -real time applications

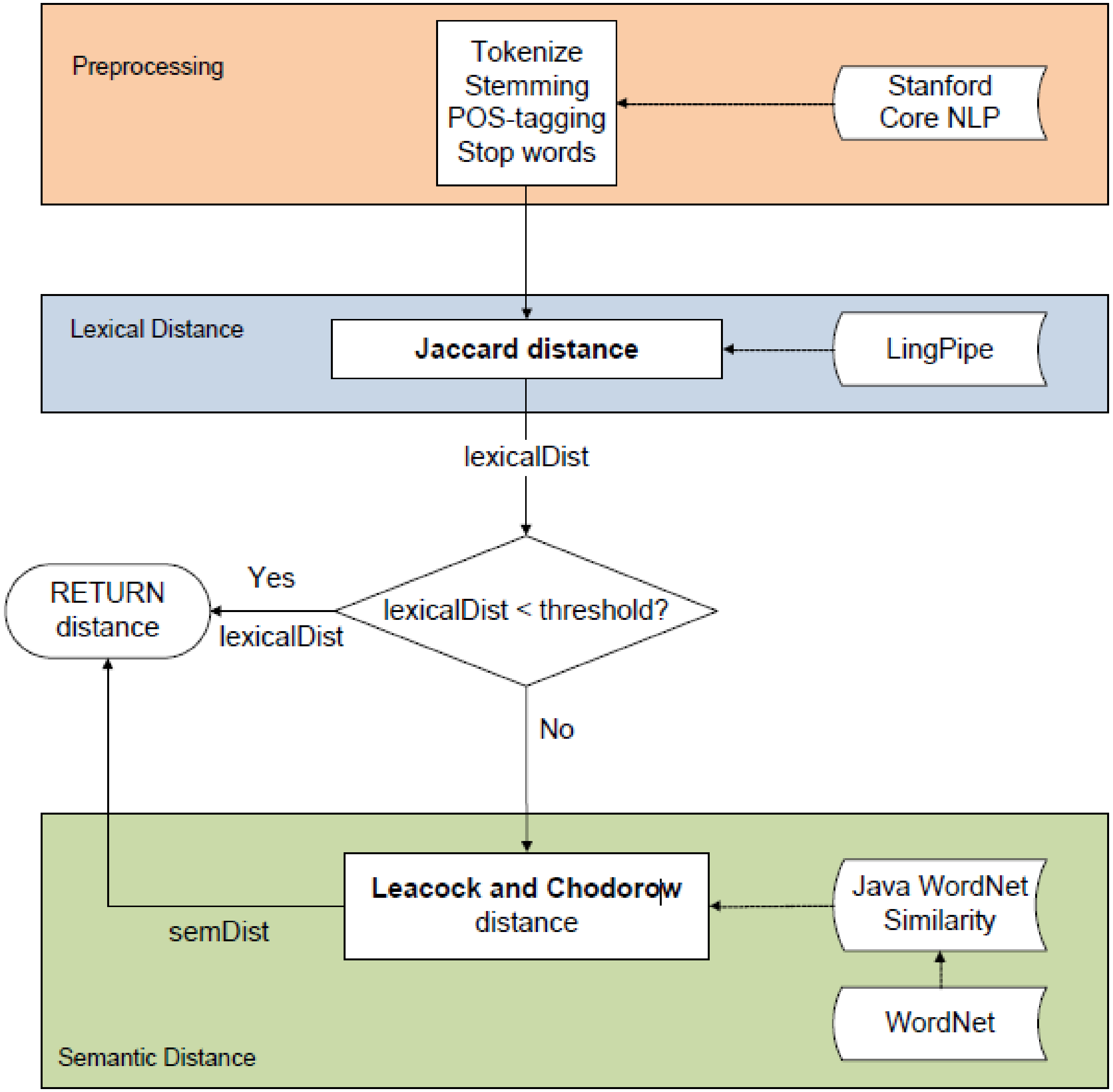
- Goals:

  - more interaction during conferences
  - better experience for the audience

Tools used

- **Stanford CoreNLP** – tokenization, POS-tagging, lemmatization
- **LingPipe** – hierarhical clustering, lexical distance
- WordNet – semantic distance

Architecture overview



Clustering

- Hierarchical clustering

pros:

- offers a structure view of the clusters involved: a tree structure called denrdogram
- by cutting the dendrogram at a certain value of the similarity different clusters can be obtained

cons:

- more costly than flat clustering

LingPipe

- { "a", "aaa", "aaaaaa", "aaaaaaaaaa" }

Complete Link

```
9.0
    4.0
        aaaaaa
        aaaaaaaaaa
    2.0
        aaa
        a
```

Single Link

```
4.0
    3.0
        aaaaaa
        2.0
            aaa
            a
    aaaaaaaaaa
```

LingPipe (2)

- Set a distance bound and maintain every cluster formed at less than or equal to that bound

```
  Set<Set<String>> clKClustering =
     clDendrogram.partitionDistance(maxDistance)
```

- Continue cutting the highest distance cluster until a specified number of clusters is obtained.

```
for (int k = 1; k <= clDendrogram.size(); ++k) {
    Set<Set<String>> clKClustering =
  clDendrogram.partitionK(k);
    System.out.println(k + "  " + slKClustering);
}
```

Question class

```
 String originalText;
String parsedText;
ArrayList<CoreLabel> allTokens;
ArrayList<CoreLabel> filteredTokens;
ArrayList<String> nouns;
ArrayList<String> verbs;
ArrayList<String> adjectives;
ArrayList<String> lemmas;
 ArrayList<String> stopVerbs;

int id;
int setId;
```

Lexical Distance class

- Ling Pipe Jaccard distance

- no_words_in_both_questions / total_number_of_unique_words

- E.g.:

Does your sister love cats? -> sister love cat
Do you love your sister -> love sister

- The Jaccard distance would be: 2/5 = 0.4 -> highly similar

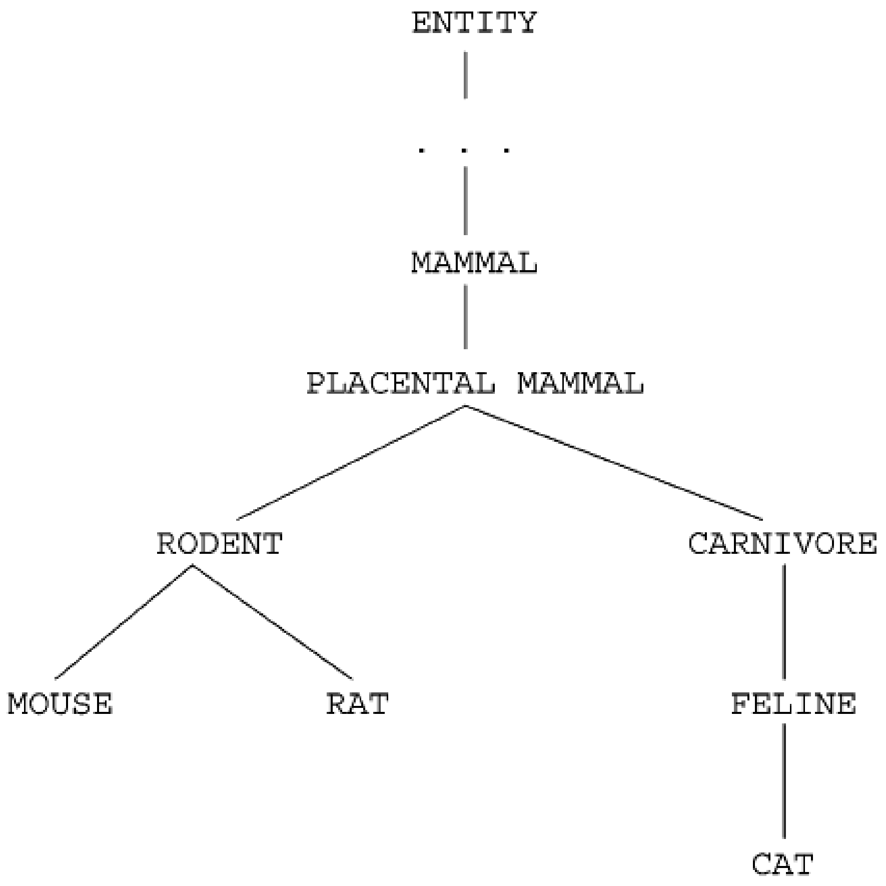
Semantic similarities

- Knowledge based semantic similarities using WordNet

- WordNet
  - Concepts represented as hierarchies
  - Each node as a synset (group of synonyms)

- Similarities based on:
  - Shortest path between two concepts (edge approach)
  - Depth of concepts (node approach)
  - Depth of least common subsumer

An example of WordNet hierarchy



SemanticDistance class

- WordNet package: Leacock & Chodorow similarity

$$Sim_{lch} = -\log \frac{length}{2 * D}$$

- Similarity between questions for nouns/verbs

$$sim_{sem} = \frac{1}{2}\left(\frac{\sum_{ai \in Q1} \max ssim\,(a_i, Q_2)}{|Q_1|} + \frac{\sum_{bj \in Q2} \max ssim\,(b_j, Q_1)}{|Q_2|}\right)$$

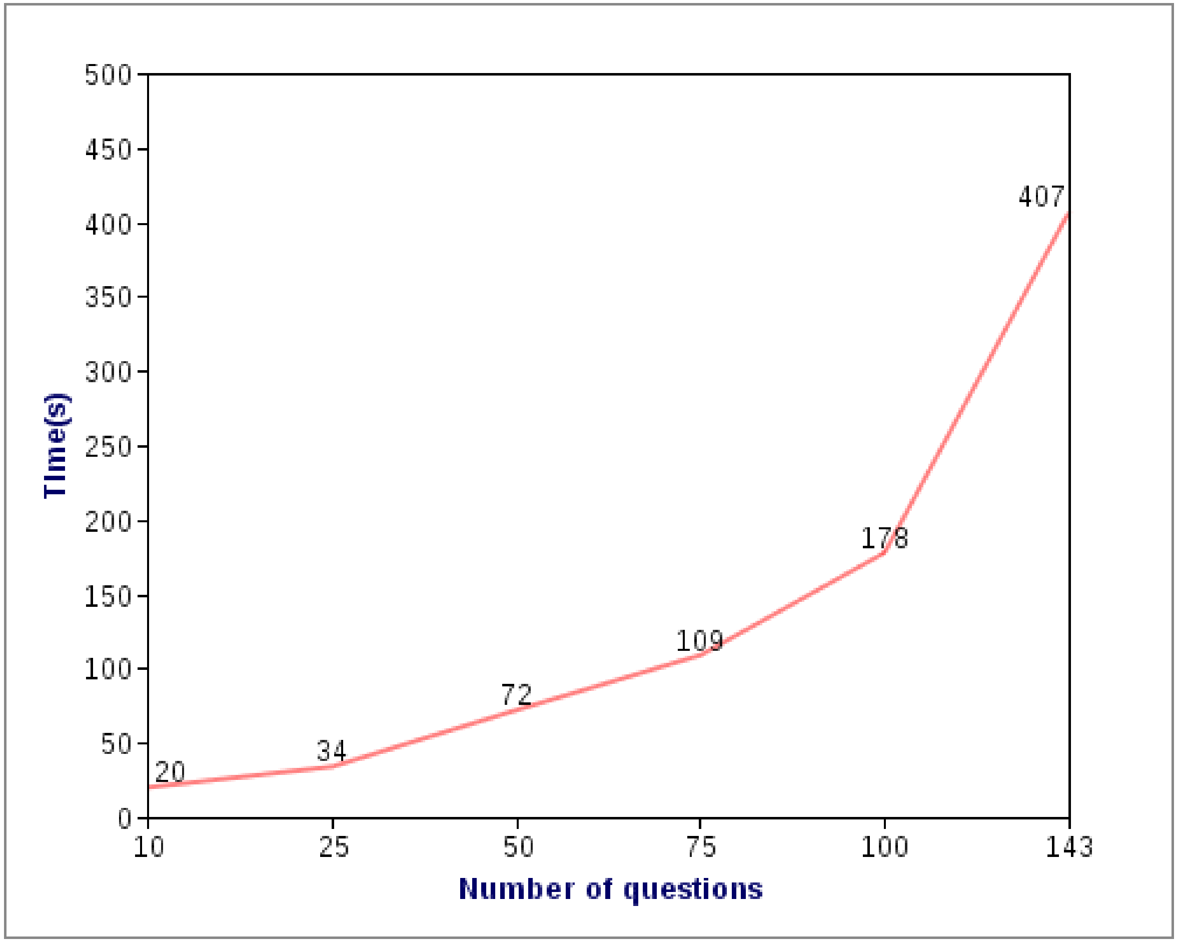
- Distance between questions

1 – **(sim_sem_of_nouns + sim_sem_of_verbs) / 2**

Evaluation (1)

```
0.7646569784324155
           0.753494911695385
              0.5
                 0.33333333333333337
                    What is the temperature of the sun 's surface?
                    The sun 's core , what is the temperature ?
                 What is the earth 's diameter ?
              Why does the moon turn orange ?
           What city had a world fair in 1900 ?
        What is Australia 's national flower ?
```

Evaluation (2)



Conclusions

- We have managed to integrate existent NLP tools in our architecture and obtain a working solution
- We need to do more testing and evaluation, specially with semantically similar sets of questions
- Integrate our solution with Smart Presentation solution

Thank you