Semantic Textual Relatedness between African and Asian Languages

CS779 (Statistical NLP)
Prof. Ashutosh Modi

The Boys
Rajarshi Dutta
Udvas Basak
Shivam Pandey

PROBLEM DESCRIPTION

Rank the pairs of sentences by their semantic relatedness in 14 languages.

1. **Supervised**:
   a. Train on the provided dataset(to be released around a month later)
   b. Can use other datasets
      i. Will have to provide impactfulness reports for each of the datasets
      ii. Perform a comparative analysis by trying out different **SOTA** models.

2. **Unsupervised**:
   a. Have to submit a system that has not been trained on using datasets that explicitly provide relatedness scores for full sentences.
   b. However, we can use unigram or bigram related datasets and systems
   c. First glance, have to create embeddings out of sentences, and then find some sort of similarity

RELEVANT DATASETS

1. **XLING -** The XLING BLI Dataset contains bilingual dictionaries for 28 language pairs. For each of the language pairs, there are 5 dictionary files: 4 training dictionaries of varying sizes (500, 1K, 3K, and 5K translation pairs) and one testing dictionary containing 2K test word pairs. https://github.com/codogogo/xling-eval Was introduced in SemEval 2017.

2. **PARANMT-50M -** This dataset essentially is generated from Neural Machine Translation or more precisely via a back-translation approach where **English** sentences are first translated to Czech via a **Seq-2-Seq** architecture and then converted back to English which creates paraphrases with the same semantics.

3. **SICK Dataset -** This consists of a crowdsourced dataset consisting of sentence pairs taken from the **FLICKR** dataset along with corresponding **relatedness score** and tones of these sentences.

4. **Quora Duplicate Question Pairs -** This dataset consists of sentences pairs from **Quora website** and associated labels highlighting whether these are semantically related or not.

PILOT DATASETS PROVIDED

1. **Big BiRD: A Large, Fine-Grained, Bigram Relatedness Dataset for Examining Semantic Composition https://aclanthology.org/N19-1050/ :** Provided for the unsupervised problem

2. Ranked Semantic Relation Dataset: Pairs of sentences with semantic relatedness scores and then ranked. Has 5500 Entries

3. Unranked Semantic Relation Dataset: Pairs of sentences, provided mostly in some Arabic Language(possibly Urdu). 100 entries.
   Also has 5 columns named Ann1, Ann2, Ann3, Ann4, Ann5. Still did not successfully understand what these mean.

PRELIMINARY IDEAS

1. Need to read up and work on Transfer Learning Techniques, will be hugely important in both the subtasks

2. A slight understanding of the semantics and structure of the other languages in question will help hugely

3. Need to plan on a single evaluation metric for each of the subtasks which helps to capture the semantic closeness efficiently

    a. For the Unsupervised Learning Task, mostly papers suggest to use simple and popular evaluations metrics like:

        i. Dice coefficient
        ii. Bhattacharyya distance
        iii. Jaccard similarity
        iv. Kendall’s tau
        v. Spearman correlation coefficient
        vi. Pearson correlation coefficient

RELATED WORK

1. **What makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study? https://arxiv.org/pdf/2110.04845.pdf (Mar 2023)**
    a. 7 data sources used. Has used similar to the 5500 entry dataset
    b. Split-half reliability as a metric.
    c. Uses Lexical Overlap(Dice Coeff), Related Words with same POS(Spacy), Related Subject and Object(Spacy SVO Extractor) and Common Words
    d. Unsupervised Static Embedding using Word2Vec, FastText, GloVe and contextual using BERT and RoBERTa.
    e. Supervised Fine-Tuning using some portion of domain-specific data from the 7 mentioned sources

2. **Just Rank: Rethinking Evaluation with Word and Sentence Similarities Bin Wang , C.-C. Jay Kuo , Haizhou Li - 2022**
    a. Introduced EvalRank
    b. Modification of Cosine and L2 Similarity
    c. Some Ranking criteria introduced, like MRR and Hits@k
    d. EvalRank approach: Focuses on local distance between points.
        i. Requires pivot sample to have longer distances to background samples than its positive candidate.
        ii. Advantages of local distance measurement:
            1. Learned embedding space forms a manifold.
            2. Approximates Euclidean space locally.
        iii. Simple similarity metrics (cosine, l2) not suitable for modeling long-distance relationships.

RELATED WORK

1. **Semantic textual similarity for modern standard and dialectal Arabic using transfer learning - Mansour Al Sulaiman, Abdullah M. MoussaID, Sherif Abdou , Hebah Elgibreen, Mohammed Faisal, Mohsen Rashwan(2022)**

    a. MSA Arabic Dataset used, also available in Egyptian and Saudi Arabic.
    b. 3 Approaches
        i. Train on SBERT based on ArabicBERT, fine-tuned using Automatic Translation of Arabic of SNLI and MultiNLI Datasets.
        ii. Interleaving English STS Dataset with ArabicBERT model using Transfer Learning
        iii. Use Knowledge Distillation based STS Models as a base and fine-tune using proposed translated dataset.
    c. Pairs of sentences in the dataset have been inputted as the Siamese Architectures.

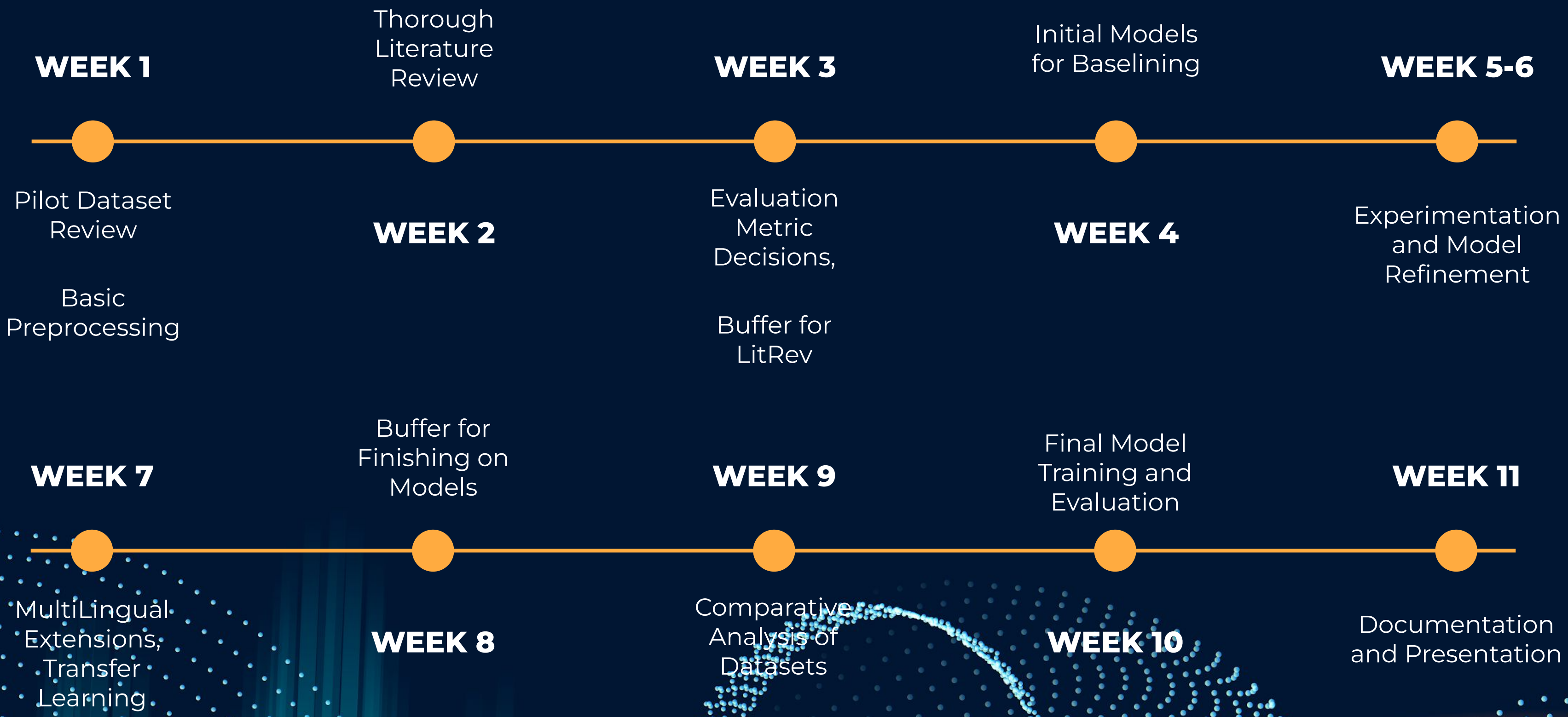
TENTATIVE WORK DISTRIBUTION

Rajarshi: Mostly look at the Datasets available, and find out what might help us in the long run. Also, do some study on the BERT and RoBERTa architecture.

Shivam: More Research Oriented, deep dive along the resources found, come up with a brief idea of the research work already done to get a better hang of the problem and the possible solutions at hand.

Udvas: Plan is to look upon each of the 14 languages structurally and semantically. Basic understanding of all the languages. Deep dive into BERT and RoBERTa.

PROPOSED TIMELINE



WEEK 1

Pilot Dataset Review

Basic Preprocessing

WEEK 2

Thorough Literature Review

WEEK 3

Evaluation Metric Decisions,

Buffer for LitRev

WEEK 4

Initial Models for Baselining

WEEK 5-6

Experimentation and Model Refinement

WEEK 7

MultiLingual Extensions, Transfer Learning

WEEK 8

Buffer for Finishing on Models

WEEK 9

Comparative Analysis of Datasets

WEEK 10

Final Model Training and Evaluation

WEEK 11

Documentation and Presentation