Project2 Proposal - Ganesh S(gs563),Kevin Li(kyl27),Remya (rb628),Yashaswini (ys429)

**Supervised Word Sense Disambiguation**

**General algorithm for processing surrounding context:**

1. Tokenize context based on words, removing punctuation.
2. Reduce each word-token to its root using WordNet or other library package.
3. Remove non-content words (i.e. a, at, and, the, of, so, …).
4. For each unique token, compute a score based on distance to target word and frequency. Score is increasing function w.r.t. frequency and distance-1. [1]
5. Generate sorted list of tokens with associated scores.
6. Keep list of tokens and scores as the value for each hashmap entry of a unique word/sense number pair.

**Trained Hashmap Structure:**

Key

* Root word
* Sense number

Value

* Sorted context token list, with scores

**General Word Sense Disambiguation Algorithm:**

For each test example to disambiguate:

1. Use general algorithm for processing surround context to generate token-score list.
2. For each hashmap entry of the target word to disambiguate:
   1. Compute score based on how closely the stored token-score list matches the generated token-score list from the test example. A higher score corresponds to a closer match. [2]
3. Return the sense number of the entry whose token-score list most closely matches the generated token-score list of the test example.

**Remarks:**

[1] We can adjust the weighting of this function to place more or less weight on distance and/or frequency. Thus, if we want to consider co-location more than co-occurrence, we would place more weight on a token’s distance to the target word than its frequency in the whole context.

[2] There are many different algorithms for comparing how closely two lists “match.” We will investigate different ways to compute this “matching” score.

**Dictionary based Word Sense Disambiguation:**

**Overall Design and Implementation**:

1. ***Pre-processing:*** Parse the test data and for each sentence, extract 4-5 adverbs/adjective/ verbs/nouns surrounding the target word. We plan to use the NLTK tools.

* Add the words to the context word set only if they are not part of the stop list (ignore pronouns, prepositions and conjunctions, punctuations)
* Perform Lemmatization with Part of Speech on the context words chosen
* To expand the scope of the context words, we are planning to use the wordnet synonyms for the target word.

1. ***Find the overlapping words:*** Consider the definitions of the target word and the definitions of the context words and find the overlapping words.

* We assign probability values based on the number of overlapping words.
* If the overlapping words are consecutive we would assign higher probability values.
* Generate bigrams/trigrams of the definitions of the context words and check if they are sub-strings of the definition(s) of the target word. This is to check if there is a consecutive overlap of words.
* Check the number of overlaps and assign the probability values to the different senses of the target word. If there are only separate overlaps and no consecutive overlaps, then the probability will be the number of overlaps.

Eg: She is here & She was here

The probability value here is 2.

If there are consecutive overlaps, then the probability value will be the square of the number of words that consecutively overlap

Eg: She is here & She is there. Here the probability value assigned will be 2\*2=4

1. ***Pick the best sense for the target:*** The definition of the target word that has the highest probability is chosen as the sense of the word.

* Construct a hash table with sense ids for target word and the probability of each sense id. The sense id with the highest probability will be chosen as the sense of the target word in the context.

**Implementation Schedule:**

|  |  |
| --- | --- |
| **Project Phase** | **Timeline** |
| Implementation of Supervised and Dictionary based Word Sense Disambiguation | March 7th - 13th |
| Implementing Extensions | March 14th - 15th |
| Integration and Testing | March 16th - 17th |
| Report submission | March 18th - 19th |