# Assignment 2-Discussion Word Sense Disambiguation

Singamsetty Sandeep, 213050064
Anmol Namdev, 213050044
Sakharam Sahadeo Gawade, 213050027

7 September 2021

# Problem Statement

- Given a sequence of words, produce the synset IDs (unique identification for a sense)

- Data SemCor

- First Technique to be used: HMM-Viterbi, Overlap based WSD

- 5-fold cross validation  (for HMM based WSD)

- See if the baselines are crossed

# First Baseline

- Most Frequent Sense (MFS)

- Accuracy 69.1%

MFS selects the most frequent sense for a word (most frequent according to a corpus)

# Second Baseline

- Wordnet First Sense (WFS)

- Accuracy 48.68%

WFS takes the first sense for a word from the WordNet from a set of senses.

# Performance report of HMM-WSD

On 5 folds Cross-Validation

- Precision: 0.501413 ± 0.000475

- Recall: 0.501413 ± 0.000475

- F1-score: 0.501413 ± 0.000475

- Crossed WFS but not MFS baseline

# Word Vector Based Overlap approach

- Accuracy:  34.60 % using Lesk Algorithm

- MFS and WFS Baseline not crossed

# Confusion Cases (for HMM)

1) In the ballroom below the dark had given way to moonlight coming in through the **BANK** of French windows

Actual: an arrangement of similar objects in a row or in tiers.(**bank.n.04**) Ex- He operated a bank of switches.

Predicted: a financial institution that accepts deposits and channels the money into lending activities.(**depository_financial_institution.n.01**) Ex- He cashed a cheque at the bank.

2) Systems measured **CIRCUIT** Resistance.

Actual: an electrical device that provides a path for electrical current to flow.(**circuit.n.01**) Ex- We need to make an LCR Circuit.

Predicted: a journey or route all the way around a particular place or area.(**tour.n.01**) Ex-They took an extended tour of Europe.

3) Lawyers **REPRESENTING** Texas Bunkers Association

Actual: be representative or typical for.(***represent.v.03***) Ex- This period is represented by Beethoven.

Predicted: express indirectly by an image, form, or model; be a symbol.(***typify.v.02***) Ex- What does Statue of Liberty Symbolize?

4) Throughout the **WORK** of the later ninth century a calm, severe serenity displays itself.

Actual: a product produced or accomplished through the effort or activity of an agency.(***work.n.02***) Ex- the work of an active imagination

Predicted: exert oneself by doing mental or physical work for a purpose or out of necessity.(***work.v.01***) Ex- I will work hard to improve my Grades.

5) The Cooking odors were **STRONGER** all over the city, at this hour, housewives would be fussing over stoves.

Actual: not faint or feeble.(***strong.s.02***) Ex- Strong odour of Burning Rubber

Predicted: having a force of authority.(***potent.s.02***) Ex- strong weapons

# Interpretation of confusion (error analysis: HMM)

- Throughout the **WORK** of the later ninth century a calm, severe serenity displays itself. Actual:***work.n.02*** Predicted:***work.v.01***

  The sense depends on the sense of previous word and probability of the word having the current sense. In the above example WORK is predicted as work because WORK occurred with sense ***work.v.01*** with a higher probability.

- Systems measured **CIRCUIT** Resistance.

  Actual:***circuit.n.01*** Predicted:***tour.n.01***

  HMM based model only refers to its previous word to tag a word. For circuit the words that help in disambiguating circuit are "systems" and "resistance", but HMM looks only at "measured" to tag "circuit".

# Confusion Cases (for Word Embedding and Overlap)

You shoot the inevitable footage of your child taking its first steps, you have merely recorded an historical event.

 1)**TAKING**

Actual: carry out (take.v.01)

Predicted: the income or profit arising from such transactions as the sale of land or other property (return.n.06)

2)**RECORDED**

Actual: make a record of; set down in permanent form (record.v.01)

Predicted: anything (such as a document or a phonograph record or a photograph) providing permanent evidence of or information about past events(record.n.01)

3) The members of our congressional delegation would like to **SEE** the resolution be passed.

Actual:perceive or be contemporaneous with(witness.v.02)

Predicted:a seat within a bishop where his cathedral is located(see.n.01)

The Jury **SAID** it did **FIND** that many of Georgia's registration and election laws are outmoded or inadequate and often ambiguous.

4) **FIND**

Actual: decide on and make a declaration about(rule.v.04)

Predict: a productive insight(discovery.n.03)

5) **SAID**

Actual: report or maintain(allege.v.01)

Predict: chance to *speak* (say.n.01)

# Interpretation of confusion (error analysis: WE-Overlap)

- You shoot the inevitable footage of your child taking its first steps, you have merely recorded an historical event.

- <u>Actual</u>: **take.v.01**  <u>Predicted</u>: *return.n.06*

  - Due to the higher number of words in the predicted sense's gloss, more word with higher similarity with context words are possible

- <u>Actual</u>: **record.v.01** <u>Predicted</u>: **record.n.01**

  - Some common context words may be present in the predicted sense's gloss which may give a false high similarity

- The Jury said it did find that many of Georgia's registration and election laws are outmoded or inadequate and often ambiguous.

- <u>Actual</u>: **allege.v.01** <u>Predicted</u>: **say.n.01**

- Some senses may vary in cosine similarity by a little(this effect may reduce if WSD is preceded by POS tagging)

# Data Processing Info (Pre-processing: HMM)

Challenges in preprocessing :

Three types of tags in semcor

1) No tag (synset ID = -1) (It is a list in corpus)

2) Tagged with a lemma of synset: A lemma is a data structure which contains

synset along with its synset_id, hypernymy, hyponymy etc.

3) Tagged with lemma but the synset ID doesn't exist in wordnet (synset ID= -2)

Stop words and words that doesn't have tag are excluded.

Probabilities are calculated similar to HMM.

Lower-cased every word of the sentence.

# Data Processing Info
# (Pre-processing: WE-Overlap)

- Existence of function words can lead to a false similarity, hence they were removed

- Words in corpus such as "preexistence" was not available as a word embedding

# WSD continuation
# Performance report of HMM-WSD

On 5 folds Cross-Validation

Following are the precision, recall and F1 values

Earlier with HMM: $0.501413 \pm 0.000475$

Improved preprocessing for HMM and added

MFS and WFS for unseen words

HMM+WFS:$0.576162 \pm 0.001714$

HMM+MFS:$0.578440 \pm 0.001010$

# WSD continuation
# Performance report of Overlap based WSD

- Accuracy: 48.38 % [Improvement using overlap with examples and their pos tags to filter out synset]
- Very close to the WFS base line

# Thank You