22201030 [illegible]

Surya kiran

25/09/23

NLP - ASSIGNMENT - 1

1. Illustrate good turing discovery.

→ Good turing Intuition:
- use the count of things which are seen once to help estimate the count of things never seen.
- Similarly, we count of things which occur 'c+1' times to estimate the count of things which occur 'c' times

- Let N_c be the no. of things that occur 'c' times. i.e frequency of 'c'.

- MLE [Maximum likelihood estimation] count for N_c is 'c' whereas, good turing estimate function of N_{c+1} is:

$$c^* = \frac{(c+1) N_{c+1}}{N_c}$$

- Re-estimates the probability of low count ngrams by considering ngrams w/ higher counts, thus provides a discounted probability value.

- Also, estimates the probability of zero count things by considering ngrams w/ count 1.

Note: N_c is the count of things we've seen 'c' times.

Let us consider an example:

ex: You are fishing and have caught the following fishes:

- carp: 10
- perch: 3
- whitefish: 2
- trout: 1
- salmon: 1
- eel: 1

Total: 18 fish. ; $N_1 = 3$; $N_2 = 1$; $N_3 = 1$; ... ; $N_{10} = 10$;

i) How likely is it that next species is salmon?

↳ $1/18$ (MLE)

ii) How likely is it that next species is a new species unlike catfish?

↳ generally, we say $0/18$ (MLE) ($c=0$), but using over intuition to estimate things we saw once to estimate the new things, we get: $N_1/N = 3/18$ ($\because N_1 = 3$)

Then, the likelihood of the next species being a salmon will be discounted, using the below expression:

↳ $c^* = \frac{(c+1)N_{c+1}}{N_c}$ → $c = 1$

MLE $p = 1/18$

→ $c^*(\text{salmon}) = \frac{(1+1)N_{1+1}}{N_1} = \frac{2N_2}{N_1}$

$= 2 \times \frac{1}{3} = 2/3$

$\therefore P^*_{GT}(\text{salmon}) = 2/3 \,/\, 18$

2. What are the issues w/ tag indeterminacy and tokenization?

The issues w/ tag indeterminacy & tokenization can be summarized as follows:

→ Tag indeterminacy:

- Ambiguity between Multiple tags:

Tag indeterminacy occurs when a word can be associated w/ multiple pos tags, and it is challenging to disambiguate which tag is correct.

- Handling indeterminate tags:

Some taggers allow the use of multiple tags for indeterminate words. In some cases, tags are placed w/ single tag or one tag is chosen during training.

- Common indeterminate tags:

Examples of common tag indeterminacies include distinguishing b/w adjectives, preterites, and past participles (eg JJ/VBD/VBN) and deciding whether a word is an adjective or noun as pronominal modifier (eg. JJ/NN).

→ Tokenization:

- Role in Word Splitting:

Tokenization is crucial for splitting words and determining words boundaries. for example, in the Penn Treebank and British National Corpus, contractions and possessive forms are separated from their stems during tokenization.

- Word Splitting Differences:

Different corpora may tokenize words differently. for instance, the Penn Treebank treats multiple part word (eg New York) as a separate words, while other taggers like CST tagger, can treat them as single words.

- Unknown Words:

Most tagging algorithms use dictionaries to assign POS tags to words. However, the creation of proper names, acronyms and new common nouns or verbs occurs regularly, and the taggers needs methods to handle the unknown words.