BPE, WordPiece Numerical Examples

Dr. Sambit Praharaj

Assistant Professor (II)

Subword Tokenization

- BPE and WordPiece
- Detailed Numerical Walkthrough

Corpus Used

- low (5)
- lower (2)
- newest (6)
- widest (3)

Why Subword Tokenization?

- Controls vocabulary size
- Handles unseen or out of vocabulary words
- Balances granularity

BPE: Initial Setup

- Split words into characters + </w>
- Initial vocab size = 11

BPE Pair Frequencies (Iter 1)

- (e,s)=9, (s,t)=9, (t,</w>)=9
- (l,o)=7, (o,w)=7, (w,e)=8

BPE Merge 1

- Merge (e,s) → es
- Vocab = 12

BPE Merge 2

- Merge (es,t) → est
- Vocab = 13

BPE Merge 3

- Merge (est,</w>) → est</w>
- Vocab = 14

BPE Merge 4 & 5

- (l,o) → lo
- (lo,w) → low
- Vocab = 16

BPE Merge 6–9

- we, nwe, newest</w>, wi
- Stop at vocab = 20

Final BPE Vocabulary

- 11 base symbols + 9 merges = 20

WordPiece: Initial Setup

- Uses continuation marker ##
- Initial vocab size = 11

Initial WordPiece Tokenization

- low → l ##o ##w
- lower → l ##o ##w ##e ##r
- newest → n ##e ##w ##e ##s ##t
- widest → w ##i ##d ##e ##s ##t

WordPiece Token Frequencies

- l=7, n=6, w=3
- ##o=7, ##w=13, ##e=17
- ##s=9, ##t=9, ##i=3, ##d=3

WordPiece Scoring Formula

- score(x,y)=count(xy)/(count(x)*count(y))

WordPiece Merge 1

- l + ##o → lo (highest score)
- Vocab = 12

WordPiece Merge 2

- ##s + ##t → ##st
- Vocab = 13

WordPiece Merge 3

- lo + ##w → low
- Vocab = 14

WordPiece Merge 4

- ##e + ##st → ##est
- Vocab = 15

WordPiece Merge 5

- w + ##i → wi
- Vocab = 16

WordPiece Merge 6

- wi + ##d → wid
- Vocab = 17

WordPiece Merge 7

- wid + ##est → widest
- Vocab = 18

WordPiece Merge 8

- n + ##e → ne
- Vocab = 19

WordPiece Merge 9 (Stop)

- ne + ##w → new
- Vocab = 20 → STOP

Final WordPiece Vocabulary

- Base tokens + 9 learned WordPieces = 20

BPE vs WordPiece

- BPE → Frequency based
- WordPiece → Likelihood based
- Both stop at vocab limit

Conclusion

- BPE favors frequent patterns
- WordPiece favors informative patterns