Concept Mapping of Medical Terms

via Syntax, Semantics, and Pragmatics Levels

Shuyue Jia
Objectives

Non-standard Terms

mapping

Standard Terms (Preferred Names)

e.g., 肺癌, 恶性肺上皮肿瘤, 肺部上皮恶性肿瘤, 肺癌, 肺恶性肿瘤(上皮来源), ......

e.g., 肺上皮恶性肿瘤
Method

Input Non-standard Term

'B神經', '梢', '末'

'BPE Algorithm for Sub-word'

'Sub-word Frequency'

'神經', False, False

'末梢'

'神經', '周围神经系统'

'周围神经系统'

'BPE Algorithm for Sub-word'

'Syntax-level
Frequentist-level
based on characters'

'末梢'

'神經'，'周围神经系统'

'神經'

'Semantics-level'

'Pragmatics-level'

'Knowledge Graph'

'Knowledge'

'Word Embedding Similarity'

Output Standard Term
Method – Syntax level Analysis

Why using Sub-words?

1. Learn **compounding** and **transliteration** from sub-word representations
2. Generalize to **translate** and **produce** new words (unseen at training time)
Getting **Sub-words:**

**Byte Pair (2-gram) Encoding (BPE) Algorithm Recap:**

1) Initialize symbol vocabulary with character vocabulary
   
   \[
   \text{vocab} = \{\text{'low': 5, 'lower': 2, 'newest': 6, 'widest': 3}\}
   \]

2) Find the most frequent 2-gram pairs ('A', 'B') from every word
   
   \[
   \{(\text{'d', 'e'}): 3, (\text{'e', 'r'}): 2, (\text{'l', 'o'}): 7, (\text{'w', '.'}): 5, (\text{'w', 'e'}): 8, (\text{'e', 'w'}): 6, (\text{'}): 2, (\text{'w', 'i'}): 3, (\text{'e', 's'}): 9, (\text{'}): 6, (\text{'s', 't'}): 9, (\text{'i', 'd'}): 3, (\text{'t', ''}): 9, (\text{'o', 'w'}): 7\}
   \]
   
   We find ('e', 's'): 9

3) Merge ('A', 'B') \(\rightarrow\) ('AB') and repeat 2).
   
   \[
   \{\text{'low': 5, 'lower': 2, 'newest': 6, 'widest': 3}\}
   \]

4) Stop merging until reach the \text{num(merge operation)} or minimum frequency
Method – Pragmatics level Analysis

Using Knowledge Graph

1. Using the **Knowledge** and **Common Sense** (languages use) for concept mapping
2. Pre-defined **Look-up Dictionary** in the field of Medical AI
Find the most similar concept via word Embedding’s Cosine Similarity

1. Two-gram
2. Sub-words
3. Forward-string grams
4. Backward-string grams
5. (If there is) Negative & other character gram
Current Challenges

1. Negative words → Add negative word’s embedding to sub-word Embedding

2. English words & Numbers → Sub-word Max Frequency

3. Punctuations

4. Abbreviations

5. Isomorphism ← Similar Word Embedding!
Thanks and have a nice day!