Behaviour of GATE Chunker and Metamap Concept Mapping in LAPPS

This document summarizes the results from testing the GATE Chunker. It also summarizes the performance of NLM's Metathesaurus (aka Metamap) using the chunker results.

Chunker Evaluation

This study categorizes different rules of the chunker by combinations of POS tags and their equivalence to Metamap concepts.

- *The NounChunker does not recognize section headings.*
  Nature (NNP), Specimen (NNP), Gross (NNP), Final (JJ) is separated into different chunks. The SNOMED code found is as follows.
  Nature - C0449499 Type of injury
  Specimen - C0438726 Sweat specimen
  Gross - C0267604 Gross' disease
  Final - C0442701 Late expiration
- *NNP NN*
  Random biopsy - no code found
  Sigmoid biopsy - C0586751 Sigmoid colon biopsy sample
  Colonic mucosa - C0227350 Lamina muscularis of colonic mucous membrane
- *NNP CD)*
  Sample 2) - no code found
- *NN + symbol*
  formalin, - C0729395 Formalin fumes
  Sigmoid Polyp" - C0586719 Sigmoid colonic polyp sample
  Blocks) - C0179321 Bite block
  Random Colon" - no code found
- *DT NN (VBZ POS)*
  the patient's - C0150831 Patient sex
  The stalk margin - no code found
  no crypt abscesses, - C0333374 Crypt abscess
  no dysplasia - C0334044 Dysplasia
  No cryptitis, - no code found
  no glandular distortion no code found
- *JJ NN*
  soft pink tissue - no code found
  tubulovillous adenoma - C0334307 Tubulovillous adenoma
- *CD*
  1.4 - C4517503 1.4

1.0 - C0588004 Baby BW
0.4 - C4517457 0.4
0.1 - C4517420 0.1
0.6 - C4068883 0.6
0.2 - C4517436 0.2
5 - C0439075 -5

- *CD NN.*
  0.1 cm. - C1269977 pT1a: Tumor more than 0.1 cm but not more than 0.5 cm in greatest dimension (breast)
  0.4 cm. - no code found
  0.9 cm. - no code found
  6 Blocks) - no code found
  5 pieces - no code found
  4 pieces - no code found
  3 pieces. - No code found
- *NNP (+symbol)*
  Sectioned - no code found
  Fragments - C0334202 Decidual fragments
  Colon- - C0009368 Colon structure (body structure)
- *NN JJ NN*
  intramucosal adenocarcinoma/high grade dysplasia. - No code found
- *JJ NN CC NN*
  high grade dysplasia and intramucosal carcinoma. - No code found
- *NNP # CD*
  Blocks #1- - C0179321 Bite block
- *DT NNP # CD*
  each Blocks #4- - no code found
  each Block #6- - no code found

Ways to improve the current chunker system:

- Add an additional section heading detector.
- Remove the word final non-alphabetic characters included by the current chunker.
- Separate the chunk if there is a chunk which includes ‘and’ or ‘/’.

Chunker Evaluation

We conclude the above ways to improve the chunker system by evaluating the GATE Chunker result by different categories of concepts. We evaluate whether the chunker returns sensible values for different categories. We investigate 17 categories from SNOMED CT Concepts.

- Administrative value

- **Body structure**

  ***Correct*:**

  soft pink tissue
  tubulovillous adenoma
  Well differentiated adenocarcinoma
  tubular adenoma
  LEFT BASE
  Benign prostatic tissue
  focal chronic inflammation
  Adenocarcinoma
  Benign prostatic glands and stroma
  RIGHT LATERAL MID
  LEFT LATERAL APEX
  Focal atypical small acinar proliferation

  ***Not Correct*:**

  Colon- – 1 chunk with bad character: Colon-
  Rectum" – 1 chunk with bad character: Rectum"
  Lymphoid aggregate – 2 chunks: a few lymphoid aggregate
  intramucosal adenocarcinoma/high grade dysplasia – 1 chunk instead of 3 : intramucosal adenocarcinoma/high grade dysplasia
  high grade dysplasia and intramucosal carcinoma – 1 chunk instead of 3 : high grade dysplasia and intramucosal carcinoma
  Signet ring type – 3 chunks instead of 1: Signet ring type
- **Clinical finding**

  ***Correct*:**

  no glandular distortion
  no dysplasia

  ***Incorrect*:**

  No cryptitis – 1 chunk with bad character: No cryptitis,
  no crypt abscesses – 1 chunk with bad character: no crypt abscesses,
  Prostate Perineural invasion absent – 2 chunks instead of 1: Prostate Perineural invasion absent
- Environment or geographical location
- Event
- Observable entity
- Organism
- Pharmaceutical / biologic product
- Physical force
- **Physical object**

  ***Correct*:**

  Colonic mucosa

  ***Incorrect*:**

one cassette – 1 chunk with bad character: one cassette.

- **Procedure**

  ***Correct*:**

  Random biopsy

  Sigmoid biopsy

  Rectum biopsy

  two soft tan core biopsies

  Incorrect:

  patient identification – 1 chunk with bad character : patient identification,

- **Qualifier value**

  ***Correct*:**

  0.1 X 0.1 X 0.1

  up to 0.4 X 0.2 X 0.1 cm.

  26 pieces

  4 pieces

  0.7 cm

  1.2 cm

  0.1 cm

  0.9 cm

  Length

  diameter

  ***Incorrect*:**

  Blocks #1 – 1 chunk with bad character: Blocks #1-

  Blocks #4 – 1 chunk with bad character: each Blocks #4-

  Blocks #6 – 1 chunk with bad character: each Blocks #6-

  5 Blocks – 2 chunks instead of 1 and bad character: 5 Blocks)

  1 Block – 2 chunks instead of 1 and bad character: 1 Block)

  0.1 X 0.1 X 0.1 – 3 chunks instead of 1 : 0.1 0.1 0.1

  0.6 X 0.2 X 0.1 cm. – 3 chunks instead of 1 : 0.6 0.2 0.1 cm.

  1.4 x 1.0 x 0.9 cm. – 3 chunks instead of 1: 1.4 1.0 0.9 cm.

  0.4 X 0.4 cm. – 2 chunks instead of 1: 0.4 0.4 cm.

  3- 5 pieces – 2 chunks instead of 1: 3- 5 pieces

  5- 4 pieces – 2 chunks instead of 1: 5- 4 pieces

  6 Blocks – 1 chunks instead of 1 and with bad character : 6 Blocks)

- Record artefact
- Situation with explicit content
- SNOMED CT Model Component
- Social context
- Special concept
- **Specimen**

  Correct:

Specimen
Incorrect:
Sample 2 – 1 chunk with bad character : Sample 2)
Sample 3 – 1 chunk with bad character: Sample 3)
Sigmoid Polyp – 1 chunk with bad character: Sigmoid Polyp”
Random Colon – 1 chunk with bad character: Random Colon”

- Staging and scales
- **Substance**
  Incorrect:
  formalin – 1 chunk with bad character: formalin,

Metamap Performance

Subsequently a post-processor was constructed to fix the chunker issue. After that, the chunks and tokens were passed into the Metathesaurus to search for the corresponding CUI code and name that matches the chunks or tokens. The matching process generally takes the following steps.

- Firstly search for chunks. If there is a CUI code related to the phrase that accept it, all tokens in the phrase will use this CUI code as feature.
  The rules deciding whether to **accept the concept** found is based on the following principles:
  - Two scores are calculated. The *term-match-score* is calculated for the proportion of words in the term matched with the words in the definition of concept. The *definition-match-score* is calculated for the proportion of words in the definition of concept matched with the words in the term searched for.
    - If the word in term matched with the word in definition except capital letter or plural, it is calculated as correctly matched.
    - If the word is a number, only treat it as correctly matched if the definition only contains the same number without any other words.
    - If the word is a single ‘x’ connecting two numbers, don’t calculate it as matched.
  - Accept the CUI concept found if: *term-match-score* is larger than 0.6 and *definition-match-score* is larger than 0.4.
- If not, search for each token in the phrase to look for the CUI code that is accepted using the same accepting principles as above.
- The CUI code related to each token is included as a feature in the BIO output file.

After implementation of this process to match the corresponding SNOMED CT concept, the performance of this matching was analysed by reviewing the percentage of correctly matched phrases or words in each category stated above from the SNOMED CT concept. The statistics were calculated over several files.

Type	SNOMED CT Concept Correct Rate
Section Heading	66/92=71.7%
Body Structure	73/90=81.1%

Qualifier Value	185/225=82.2%
Specimen	23/38=60.5%
Situation	35/72=48.6%
Clinical Finding	20/36=55.5%
Procedure	13/23=56.5%
Physical Object	2/7=28.6%
Substance	18/18=100%
Observable Entity	3/21=14.3%
Overall	438/622=70.4%