The SPECIALIST NLP Tools
Generating Multiwords from MEDLINE
Filters & Matchers

By: Dr. Chris J. Lu

The Lexical Systems Group
NLM – LHNCBC - CGSB

June, 2016

• The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov
Table of Contents

- Introduction - LexMultiwords
- Objective and Approach
  - Filters
  - Matchers
  - The Distilled MEDLINE N-gram Set
- Practice Results and Future Work
- Questions
What Is a Word?

- **Orthographic** (space in written text vs. but not in speech)
  - Single words vs. multiwords:
    - [ice-cream] vs. [ice cream]
- **Lexical Item** (lexeme) – Lexical Record in Lexicon
  - Part of Speech:
    - saw - [noun|singular|E0054443], [verb|infinitive|E0054444]
  - Inflections (grammatical word-forms):
    - 0023681|noun - [dog|singular] vs. [dogs|plural]
  - A special unit of meaning:
    - bank|E0011894 – [financial institution] vs. [margin of a watercourse]
  - Spelling Variants:
    - [color] vs. [colour], [labeled] vs. [labelled]
- Spelling
- ...

...
Single Words vs. Multiwords

- Words include single words and multiwords
- Word boundary – space or tab
- Multiwords are words that happen to be spelled with a space
- Single words vs. multiwords
  - One word can be represented as a single word or multiword (clubfoot)

<table>
<thead>
<tr>
<th>Single words</th>
<th>Multiwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>saw</td>
<td>club foot</td>
</tr>
<tr>
<td>ice-cream</td>
<td>ice cream</td>
</tr>
<tr>
<td>clubfoot</td>
<td>drop-foot gait</td>
</tr>
<tr>
<td>club-foot</td>
<td>Horner’s syndrome</td>
</tr>
</tbody>
</table>
Lexicon Unigram Coverage – Word Count

- Total word count for MEDLINE (2016): 3,114,617,940
- Lexicon covers > 98% unigrams from MEDLINE

<table>
<thead>
<tr>
<th>Types</th>
<th>Word Count</th>
<th>Percentage %</th>
<th>Accu. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEXICON</td>
<td>2,911,156,308</td>
<td>93.4675%</td>
<td>93.4675%</td>
</tr>
<tr>
<td>NUMBER</td>
<td>8,753,120</td>
<td>0.2810%</td>
<td>93.7485%</td>
</tr>
<tr>
<td>DIGIT</td>
<td>145,548,882</td>
<td>4.6731%</td>
<td>98.4216%</td>
</tr>
<tr>
<td>MULTIWORD</td>
<td>19,148,557</td>
<td>0.6148%</td>
<td>99.0364%</td>
</tr>
<tr>
<td>NEW</td>
<td>30,011,073</td>
<td>0.9636%</td>
<td>100.0000%</td>
</tr>
<tr>
<td>Total</td>
<td>3,114,617,940</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lexicon Unigram Coverage - Unique Word

- Total unique word for MEDLINE (2016): 3,619,854
- Lexicon covers 10.62% unigrams in MEDLINE

<table>
<thead>
<tr>
<th>Types</th>
<th>Word Count</th>
<th>Percentage %</th>
<th>Accu. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEXICON (S)</td>
<td>296,747</td>
<td>8.1978%</td>
<td>8.1978%</td>
</tr>
<tr>
<td>NUMBER</td>
<td>62</td>
<td>0.0017%</td>
<td>8.1995%</td>
</tr>
<tr>
<td>DIGIT</td>
<td>87,437</td>
<td>2.4155%</td>
<td>10.6150%</td>
</tr>
<tr>
<td>MULTIWORD</td>
<td>43,811</td>
<td>1.2103%</td>
<td>11.8253%</td>
</tr>
<tr>
<td>NEW</td>
<td>3,191,797</td>
<td>88.1747%</td>
<td>100.0000%</td>
</tr>
<tr>
<td>Total</td>
<td>3,619,854</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• The frequency spectrum of Alice in Wonderland, Word Frequency Distributions by R. Harald Baayen, 2001, Springer-Science + Business Media, B.V., P:10
Word Frequency vs. Rank - MEDLINE 2016
Lexicon Growth – 2002 to 2016

- 491,639 lexical records
- 1,090,050 words (categories and inflections)
- 915,583 forms (spelling only)
  - Single words: 468,655 (51.19%); Multiwords: 446,928 (48.81%)
Future Lexicon Building

- Lexicon single words: high coverage
- Lexicon multiwords (LMWs): increasing growth
- Multiwords acquisition is the key for future Lexicon building
Multiword Expression (MWE)

- Multiwords (MWEs) are used extensively in many specialized domains, particularly in areas like biomedical, medicine, computer science and engineering.
- MWEs are hard to deal with in NLP tasks:
  - have a large amount of distinct phenomena
  - lack of syntactic theories and semantic formalisms
  - phrasal preposition (because of, due to)
  - adverbs (on time)
- Non-decomposable MWEs:
  - fixed phrases (kingdom come, by and large, etc.)
  - idioms (kick the bucket, shoot the breeze, etc.)
- Utilize facts (instead of rules) to resolve the issues.
Multiwords Issues - Examples

- **Query Expansion**

<table>
<thead>
<tr>
<th>Synonym-key</th>
<th>Synonym-value</th>
<th>Query Expansion Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>perforated</td>
<td>perforation</td>
<td>perforated ear drum =&gt; <strong>perforation</strong> ear drum (Tympanic Membrane Perforation)</td>
</tr>
<tr>
<td>hot</td>
<td>warn</td>
<td>hot dog =&gt; <strong>warm</strong> dog</td>
</tr>
<tr>
<td>dog</td>
<td>canine</td>
<td>hot dog =&gt; hot <strong>canine</strong></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- The concept associated with a sentence often coincides with the longest multiword in the sentences (used in MetaMap)
Multiwords in NLP

- Identify multiwords as phrasal units directly
- Reduce part-of-speech ambiguity
- Improve stemming and lemmatization
- Better concept mapping results
The SPECIALIST NLP Tools

LexBuild Process (Computer-aided)

Source:
• Element Words
• Find multiword candidates contain an element word by Essie search engine

Review:
• Google Scholar
• Dictionaries
• Biomedical publications
• Domain-specific databases
• Nomenclature guidelines
• books
• ...

Check:
• LexAccess

Build:
• LexBuild


Issues of Element Word Approach

- Time consuming
- Essie search engine is not current (MEDLINE, 2007)
- Frequency of new words in Lexicon:
  - Use new element words (frequency rank: 1565 ~ 2549)
  - Frequency of element words (not multiwords)
  - Low frequency element words vs. high frequency multiword?
- New multiwords from old element words are missing
New LexBuild Process

Source (MEDLINE):
- LMW Candidates

Review:
- Google Scholar
- Dictionaries
- Biomedical publications
- Domain-specific databases
- Nomenclature guidelines
- books
- ...

Build:
- LexBuild
Project Objective

➢ A systematic way to add multiwords form MEDLINE to the SPECIALIST Lexicon:
  • Covers multiwords from the latest MEDLINE
  • Generates high precision multiword candidate list
    o To save time for linguists to build Lexicon
LexMultiword vs. Multiword Expression

- LMWs are a subset of MWEs

- Collocation (frequency)
  - An arbitrary statistically significant association between co-occurring items
  - “undergoing cardiac surgery” vs. “cardiac surgery”
  - “in the house” vs. “in house”

- Embedded lexical information
  - Verb particle construction (handled by complementation types)
    “beat someone up” => beat|E0012175, tran=np;part(up)
  - Light verb (information is in the lexical records, but they are not LMWs)
    “give birth”, “make love”, etc.

- Non-decomposable idioms (beyond the score of the Lexicon)
  - “kick the bucket”, “shoot the breeze”, etc.

- Design goal is set to five-grams to reach coverage above 99%
  - Most MWE research only focus on bi-grams or tri-grams
N-gram Approach

- **Source:** get all n-grams from MEDLINE documents
  - No MEDLINE n-gram set available for public
- **Matcher:** retrieve word candidates by patterns, rules, etc.
  - Inclusive filter (matcher): focus only on precision
- **Filter:** filter out n-grams that are invalid words
  - Exclusive filter: focus on not to drop recall, and then increase precision
- **Validation & Build:** Expert’s review
  - Very expensive, minimize manual process

- To bridge the gap between n-grams (statistical co-occurrence) and our term-based Lexicon.
LMWs – Processes

Filters

- (Distilled) MEDLINE
- MEDLINE N-gram Set
- n-grams not in Lexicon

Matchers

- (ACR) Pattern
- EndWord Pattern
- CUI Pattern
- SpVar Pattern

MEDLINE

MEDLINE N-gram Set

LexMultiWord Candidates
**N-gram**

- An \( n \)-gram is a contiguous sequence of \( n \) items from a given sequence of text or speech
  - An \( n \)-gram of size 1 is referred to as a "unigram"
  - Size 2 is a "bigram" (or a "digram");
  - Size 3 is a "trigram".
  - Larger sizes are sometimes referred to by the value of \( n \), e.g., "four-gram", "five-gram", and so on.

**Example:**
- to be or not to be

<table>
<thead>
<tr>
<th>( N = 1 )</th>
<th>Unigram</th>
<th>to, be, or, not, to, be</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N = 2 )</td>
<td>Bigram</td>
<td>to be, be or, or not, not to, to be</td>
</tr>
<tr>
<td>( N = 3 )</td>
<td>Trigram</td>
<td>to be or, be or not, or not to, not to be</td>
</tr>
</tbody>
</table>
N-gram Requirements*

- Range of N:

<table>
<thead>
<tr>
<th>N</th>
<th>WC</th>
<th>Accumulated WC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>468,655 (51.1865%)</td>
<td>468,655 (51.1865%)</td>
</tr>
<tr>
<td>2</td>
<td>294,022 (32.1131%)</td>
<td>762,677 (83.2996%)</td>
</tr>
<tr>
<td>3</td>
<td>102,746 (11.2219%)</td>
<td>865,423 (94.5215%)</td>
</tr>
<tr>
<td>4</td>
<td>34,339 (3.7505%)</td>
<td>899,762 (98.2720%)</td>
</tr>
<tr>
<td>5</td>
<td>10,162 (1.1099%)</td>
<td>909,924 (99.3819%)</td>
</tr>
<tr>
<td>6</td>
<td>3,483 (0.3804%)</td>
<td>913,407 (99.7923%)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>


* “Generating the MEDLINE N-Gam Set”,
  Lu, Chris J.; Tormey, Destinee; McCreedy, Lynn; and Browne, Allen C.,
  AMIA 2015 Annual Symposium, San Francisco, CA, November 14-18, 2015, P1569
# The MEDLINE N-gram Set - Specifications

<table>
<thead>
<tr>
<th>N-grams</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDLINE files</td>
<td>1-746</td>
<td>1-779</td>
<td>1-812</td>
</tr>
<tr>
<td>Max. length</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Min. WC</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Min. DC</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total documents</td>
<td>22,356,869</td>
<td>23,343,329</td>
<td>24,358,442</td>
</tr>
<tr>
<td>Total sentences</td>
<td>126,612,705</td>
<td>134,834,507</td>
<td>143,471,776</td>
</tr>
<tr>
<td>Total tokens</td>
<td>2,610,209,406</td>
<td>2,786,085,158</td>
<td>2,971,013,236</td>
</tr>
</tbody>
</table>
The MEDLINE N-gram Set


<table>
<thead>
<tr>
<th>N-grams</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>unigrams</td>
<td>804,382</td>
<td>843,206</td>
<td>883,287</td>
</tr>
<tr>
<td>bigrams</td>
<td>4,587,349</td>
<td>4,845,965</td>
<td>5,114,547</td>
</tr>
<tr>
<td>trigrams</td>
<td>6,287,536</td>
<td>6,702,194</td>
<td>7,134,807</td>
</tr>
<tr>
<td>four-grams</td>
<td>3,799,377</td>
<td>4,082,612</td>
<td>4,380,474</td>
</tr>
<tr>
<td>five-grams</td>
<td>1,545,175</td>
<td>1,674,715</td>
<td>1,812,223</td>
</tr>
<tr>
<td>n-gram set</td>
<td>17,023,819</td>
<td>18,148,692</td>
<td>19,325,338</td>
</tr>
</tbody>
</table>
Enhanced N-gram Set?

- 17 ~ 19 million is a big number (Big Data)
- Reduce the size by filtering out invalid multiwords:
  - increase precision
  - without sacrificing recall
  - distilled MEDLINE n-gram set
Filter

Filter efficiency = trap terms / total terms
Filter passing rate = pass-through terms / total terms
Good filters have high efficiency and accuracy
Accuracy Test: apply filters on Lexicon (valid word set)
- Accuracy = TP + TN / TP + TN + FP + FN
  = TP / TP + FN ….. if TN & FP are 0
  = pass / total terms
  = passing rate

<table>
<thead>
<tr>
<th></th>
<th>Trap (not retrieved)</th>
<th>Pass (retrieved)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid (relevant)</td>
<td>FN</td>
<td>TP</td>
</tr>
<tr>
<td>Invalid (not relevant)</td>
<td>TN</td>
<td>FP</td>
</tr>
</tbody>
</table>
A distilled n-gram set by filtering out invalid words.

Applied high accuracy filter \((V_0 = V_1 = \ldots = V_n; I_0 > I_1 > \ldots > I_n)\)

Higher precision with same recall rate (if filter has high accuracy rate)

N-gram Precision \(n = \frac{V_n}{V_n + I_n}\)

\[= \frac{V_0}{V_0 + I_0} \quad \ldots \quad V_n \text{ is same as } V_0 \text{ (high accuracy)}\]

\[> \frac{V_0}{V_0 + I_0} \quad \ldots \quad I_0 \text{ is bigger than } I_n \text{ (high efficiency)}\]

N-gram Recall \(n = \frac{V_n}{V_n + FN_n}\)

\[= \frac{V_n}{V_n + FN_0} \quad \ldots \quad FN_n \text{ is a constant (0), same as } FN_0\]

\[= \frac{V_0}{V_0 + FN_0} \quad \ldots \quad V_n \text{ is same as } V_0 \text{ (high accuracy)}\]
Distilled N-gram Set

MEDLINE N-gram Set

General Filters (5)

Pattern Filters (6)

Lead-End-Term Filters (5)

Domain Filters

Distilled MEDLINE N-gram Set

Domain MEDLINE N-gram Corpus
# General Exclusive Filters

<table>
<thead>
<tr>
<th>Filter</th>
<th>Accuracy On Lexicon (875,890)</th>
<th>Passing Rate N-gram Set</th>
<th>Accumulated Passing Rate</th>
<th>Trapped Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pipe</strong></td>
<td>100.0000% (0)</td>
<td>100.0000% (6)</td>
<td>100.0000%</td>
<td>• 38</td>
</tr>
</tbody>
</table>
  • 33|37|Ag|AgCl |
| **Punctuation or space** | 100.0000% (0)                  | 99.9977% (386)          | 99.9977%                 | • 1259147|3690494|=  
  • 604567|2377864|+-/ |
| **Digit**            | 99.9999% (1)                   | 99.3141% (116,772)      | 99.3118%                 | • 1404799|2062240|2  
  • 239725|499064|95% |
| **Number**           | 99.9953% (41)                  | 99.9760% (4,056)        | 99.2879%                 | • 2463066|3359594|two  
  • 18246|20674|first and second |
| **Digit and stopword** | 99.9993% (6)                   | 99.1595% (142,067)      | 98.4534%                 | • 3155416|4125616|on the  
  • 11180|12722|1, 2, and |
# Pattern Exclusive Filters

<table>
<thead>
<tr>
<th>Filter</th>
<th>Accuracy On Lexicon (875,890)</th>
<th>Passing Rate N-gram Set</th>
<th>Accumulated Passing Rate</th>
<th>Trapped Examples</th>
</tr>
</thead>
</table>
| Parenthetic acronym - (ACR)  | 100.0000% (0)                  | 99.0232% (163,714)       | 97.4917%                 | • 33117|33381|chain reaction (PCR)  
• 30095|30315|polymerase chain reaction (PCR) |
| Indefinite article           | 99.9985% (13)                  | 98.1703% (303,679)       | 95.7079%                 | • 270384|292590|a case  
• 40271|40512|A series |
| UPPERCASE colon              | 99.9999% (1)                   | 99.4302% (92,841)        | 95.1625%                 | • 2069343|2070116|RESULTS:  
• 18015|18016|AIM: The |
| Disallowed punctuation       | 99.9978% (19)                  | 99.3020% (113,073)       | 94.4983%                 | • 324405|719011|(n =  
• 86525|133350|(P < 0.05) |
| Measurement                  | 99.9967% (29)                  | 98.1947% (290,421)       | 92.7924%                 | • 154905|181001|two groups  
• 12160|15197|10 mg/kg |
| Incomplete                   | 99.9999% (1)                   | 97.8470% (340,109)       | 90.7945%                 | • 482021|1107869|(P  
• 25347|25992|years) with |
<table>
<thead>
<tr>
<th>Filter</th>
<th>Accuracy On Lexicon (875,890)</th>
<th>Passing Rate N-gram set</th>
<th>Accumulated Passing Rate</th>
<th>Trapped Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute Invalid Lead-Term</td>
<td>99.9947% (46)</td>
<td>73.0945% (4,158,702)</td>
<td>66.3658%</td>
<td>• 2780043</td>
</tr>
<tr>
<td>Absolute Invalid End-Term</td>
<td>99.997% (3)</td>
<td>78.8984% (2,384,059)</td>
<td>52.3615%</td>
<td>• 1878109</td>
</tr>
<tr>
<td>Lead-End-Term</td>
<td>99.9992% (7)</td>
<td>99.9741% (2,312)</td>
<td>52.3480%</td>
<td>• 2578756</td>
</tr>
<tr>
<td>Lead-Term no SpVar</td>
<td>99.9887% (99)</td>
<td>85.6678% (1,277,229)</td>
<td>44.8454%</td>
<td>• 658430</td>
</tr>
<tr>
<td>End-Term no SpVar</td>
<td>99.9975% (22)</td>
<td>83.1945% (1,283,001)</td>
<td>37.3089%</td>
<td>• 1009451</td>
</tr>
</tbody>
</table>
The Distilled MEDLINE N-gram Set


<table>
<thead>
<tr>
<th>N-grams</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>unigrams</td>
<td>804,382</td>
<td>843,206</td>
<td>883,287</td>
</tr>
<tr>
<td>bigrams</td>
<td>4,587,349</td>
<td>4,845,965</td>
<td>5,114,547</td>
</tr>
<tr>
<td>trigrams</td>
<td>6,287,536</td>
<td>6,702,194</td>
<td>7,134,807</td>
</tr>
<tr>
<td>four-grams</td>
<td>3,799,377</td>
<td>4,082,612</td>
<td>4,380,474</td>
</tr>
<tr>
<td>five-grams</td>
<td>1,545,175</td>
<td>1,674,715</td>
<td>1,812,223</td>
</tr>
<tr>
<td>N-gram Set</td>
<td>17,023,819</td>
<td>18,148,692</td>
<td>19,325,338</td>
</tr>
<tr>
<td>Distilled N-gram Set</td>
<td>6,351,392</td>
<td>6,793,561</td>
<td>7,402,848</td>
</tr>
<tr>
<td>Passing Rate</td>
<td>37.31%</td>
<td>37.43%</td>
<td>38.30%</td>
</tr>
</tbody>
</table>
Matcher

Filtered N-gram Set → Trapped Valid LMWs

<table>
<thead>
<tr>
<th></th>
<th>Trap (retrieved)</th>
<th>Pass (not retrieved)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid (relevant)</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Invalid (not relevant)</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

- Parenthetic Acronym Pattern Matcher
  - “computed tomography (CT)”,”magnetic resonance imaging (MRI)”, etc.

- Spelling Variant Pattern Matcher
  - Applied algorithm of SpVarNorm, Metaphone, edit distance, sorted distance, etc.

- Metathesaurus CUI Pattern Matcher
  - LMW candidate if a term has CUI(s)
  - Apply STMT to retrieve CUIs (2 subterm substitutions by their synonyms)

- EndWord pattern Matcher
  - syndrome: “migraine syndrome”, “contiguous gene syndrome”, etc.
  - disease: “Fabry disease”, “Devic disease”, etc.
### Practice Results

- **Baseline:** 16,675 LMW Candidates from (ACR) matcher, tagged by linguists

<table>
<thead>
<tr>
<th>Case</th>
<th>Test Case - Model</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Parenthetic Acronym - gold standard</td>
<td>14,805</td>
<td>1,870</td>
<td>0</td>
<td>0</td>
<td>0.8879</td>
<td>1.0000</td>
<td>0.9406</td>
<td>0.8879</td>
</tr>
<tr>
<td>2</td>
<td>Distilled MEDLINE N-gram Set (16 filters)</td>
<td>14,796</td>
<td>1,305</td>
<td>9</td>
<td>565</td>
<td>0.9189</td>
<td>0.9994</td>
<td>0.9575</td>
<td>0.9212</td>
</tr>
<tr>
<td>3</td>
<td>Spelling Variant Pattern matcher</td>
<td>7,509</td>
<td>482</td>
<td>7,296</td>
<td>1,388</td>
<td>0.9397</td>
<td>0.5072</td>
<td>0.6588</td>
<td>0.5336</td>
</tr>
<tr>
<td>4</td>
<td>Metathesaurus CUI Pattern matcher</td>
<td>9,488</td>
<td>752</td>
<td>5,317</td>
<td>1,118</td>
<td>0.9266</td>
<td>0.6409</td>
<td>0.7577</td>
<td>0.6360</td>
</tr>
<tr>
<td>5</td>
<td>EndWord Pattern matcher</td>
<td>1,710</td>
<td>180</td>
<td>13,095</td>
<td>1,690</td>
<td>0.9048</td>
<td>0.1155</td>
<td>0.2049</td>
<td>0.2039</td>
</tr>
<tr>
<td>6</td>
<td>Distilled + SpVar + CUI</td>
<td>5,510</td>
<td>206</td>
<td>9,295</td>
<td>1,664</td>
<td>0.9640</td>
<td>0.3722</td>
<td>0.5370</td>
<td>0.4302</td>
</tr>
<tr>
<td>7</td>
<td>Distilled + SpVar + CUI + EndWord</td>
<td>727</td>
<td>11</td>
<td>14,078</td>
<td>1,859</td>
<td>0.9851</td>
<td>0.0491</td>
<td>0.0935</td>
<td>0.1551</td>
</tr>
</tbody>
</table>
Frequency Analysis – Valid Words Distribution
Frequency Analysis – PRF for AEP Model

[Graph showing Precision, Recall, and F1 scores against Word Count Class]
Frequency Analysis Summary

Observation
- Most words are in the low WC range (LMWs or single words)
- N-gram in low WC range have higher normalized recall and F1 score, with precision above 0.8.
- N-grams in high WC range have very few valid LMWs, with precision between 0 and 1.

Strategy
- Set on the lower WC range (100-10,000) for multiwords
- Set on the high WC range for single words (most unigrams are valid single words)
- Applied with filters and matchers to generate LMW candidates from the MEDLINE n-gram set
Summary

- All filters have accuracy rate above 99.99% (tested on Lexicon)
- Obtain the distilled MEDLINE n-gram set at passing rate of 37-38%
  - smaller data set
  - better precision
  - similar recall
  ⇒ The recall rate between the Lexicon test set (0.9997) and baseline (0.9994) are almost identical
  - used as baseline for further analysis
- Improve lexBuilding
- Distribute the MEDLINE n-gram set (2014+) to public
- Distribute the Distilled MEDLINE n-gram set (2014+) to public
- LexBuilding on multiwords
Future Work

- Continuously enhance filters and matchers for LexBuilding on multiwords
  - Enhance SpVar Matcher model on SpVarNorm + M2CES models
  - Apply frequency strategy

- Apply different matchers to the Distilled MEDLINE n-gram set to generate LMW candidates

- Develop new SPECAILIST NLP Tools based on multiwords
Questions