

The SPECIALIST NLP Tools

Generating Multiwords from MEDLINE Filters & Matchers

By: Dr. Chris J. Lu

[The Lexical Systems Group](#)

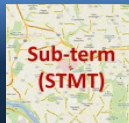
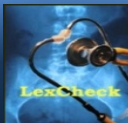
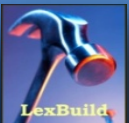
[NLM](#) – [LHNCBC](#) - [CGSB](#)

June, 2016

- Lexical Systems Group: <http://umlslex.nlm.nih.gov>
- The SPECIALIST NLP Tools: <http://specialist.nlm.nih.gov>

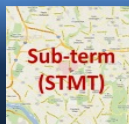
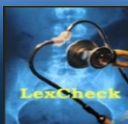
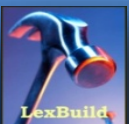
Table of Contents

- Introduction - LexMultiwords
- Objective and Approach
 - Filters
 - Matchers
 - The Distilled MEDLINE N-gram Set
- Practice Results and Future Work
- Questions



What Is a Word?

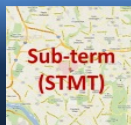
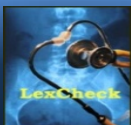
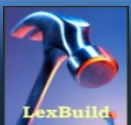
- Orthographic (space in written text vs. but not in speech)
 - Single words vs. multiwords:
 - [ice-cream] vs. [ice cream]
- Lexical Item (lexeme) – Lexical Record in Lexicon
 - Part of Speech:
 - saw - [noun | singular | E0054443], [verb | infinitive | E0054444]
 - Inflections (grammatical word-forms):
 - 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)

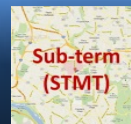
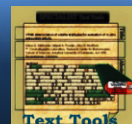
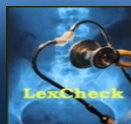
Single words	Multiwords
saw	club foot
ice-cream	ice cream
clubfoot	drop-foot gait
club-foot	Horner's syndrome



Lexicon Unigram Coverage – Word Count

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

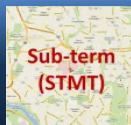
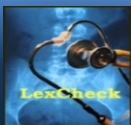
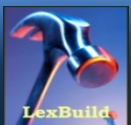
Types	Word Count	Percentage %	Accu. %
LEXICON	2,911,156,308	93.4675%	93.4675%
NUMBER	8,753,120	0.2810%	93.7485%
DIGIT	145,548,882	4.6731%	98.4216%
MULTIWORD	19,148,557	0.6148%	99.0364%
NEW	30,011,073	0.9636%	100.0000%
Total	3,114,617,940		



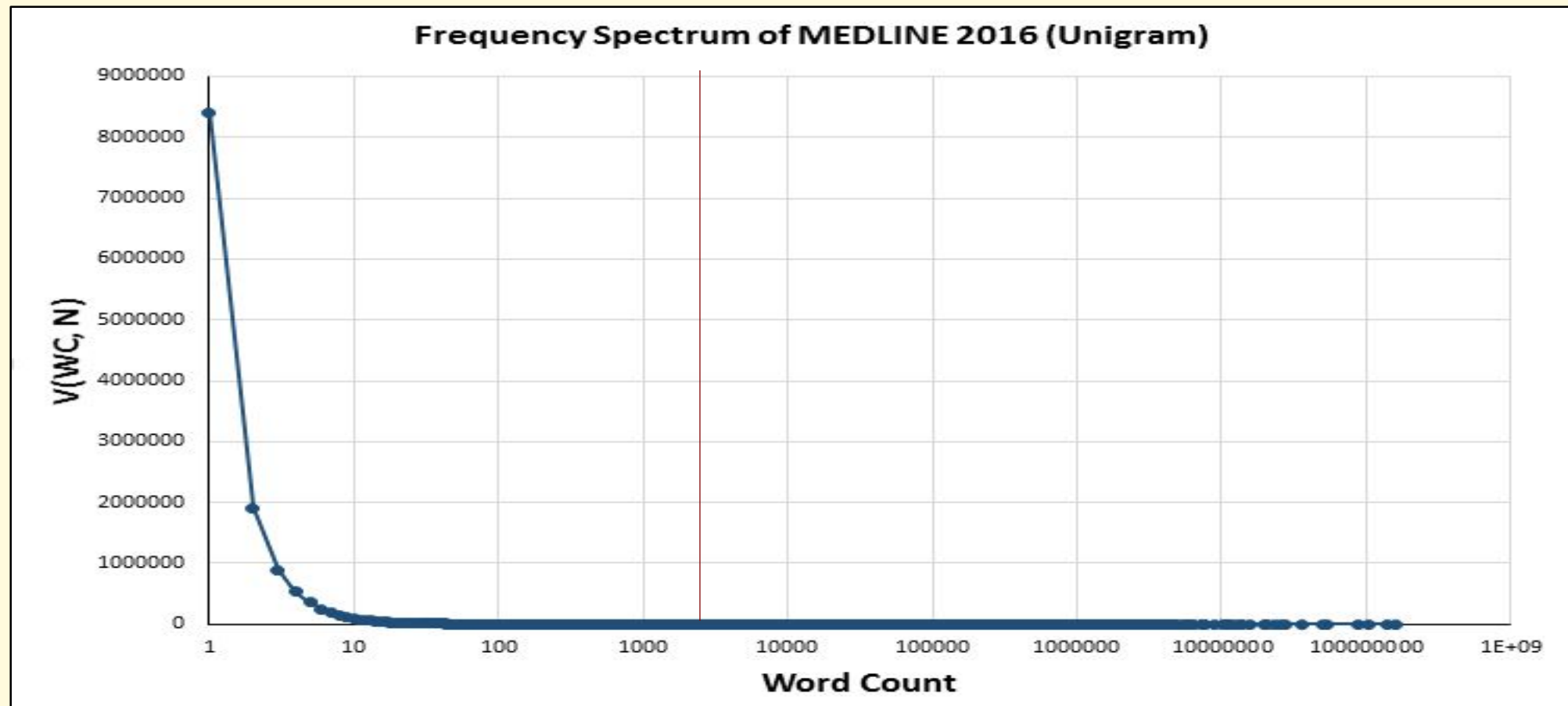
Lexicon Unigram Coverage - Unique Word

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

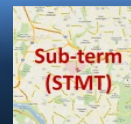
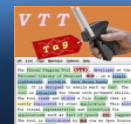
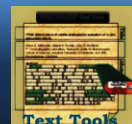
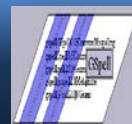
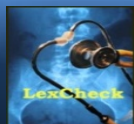
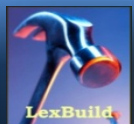
Types	Word Count	Percentage %	Accu. %
LEXICON (S)	296,747	8.1978%	8.1978%
NUMBER	62	0.0017%	8.1995%
DIGIT	87,437	2.4155%	10.6150%
MULTIWORD	43,811	1.2103%	11.8253%
NEW	3,191,797	88.1747%	100.0000%
Total	3,619,854		



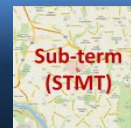
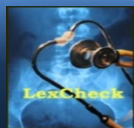
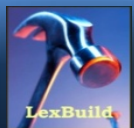
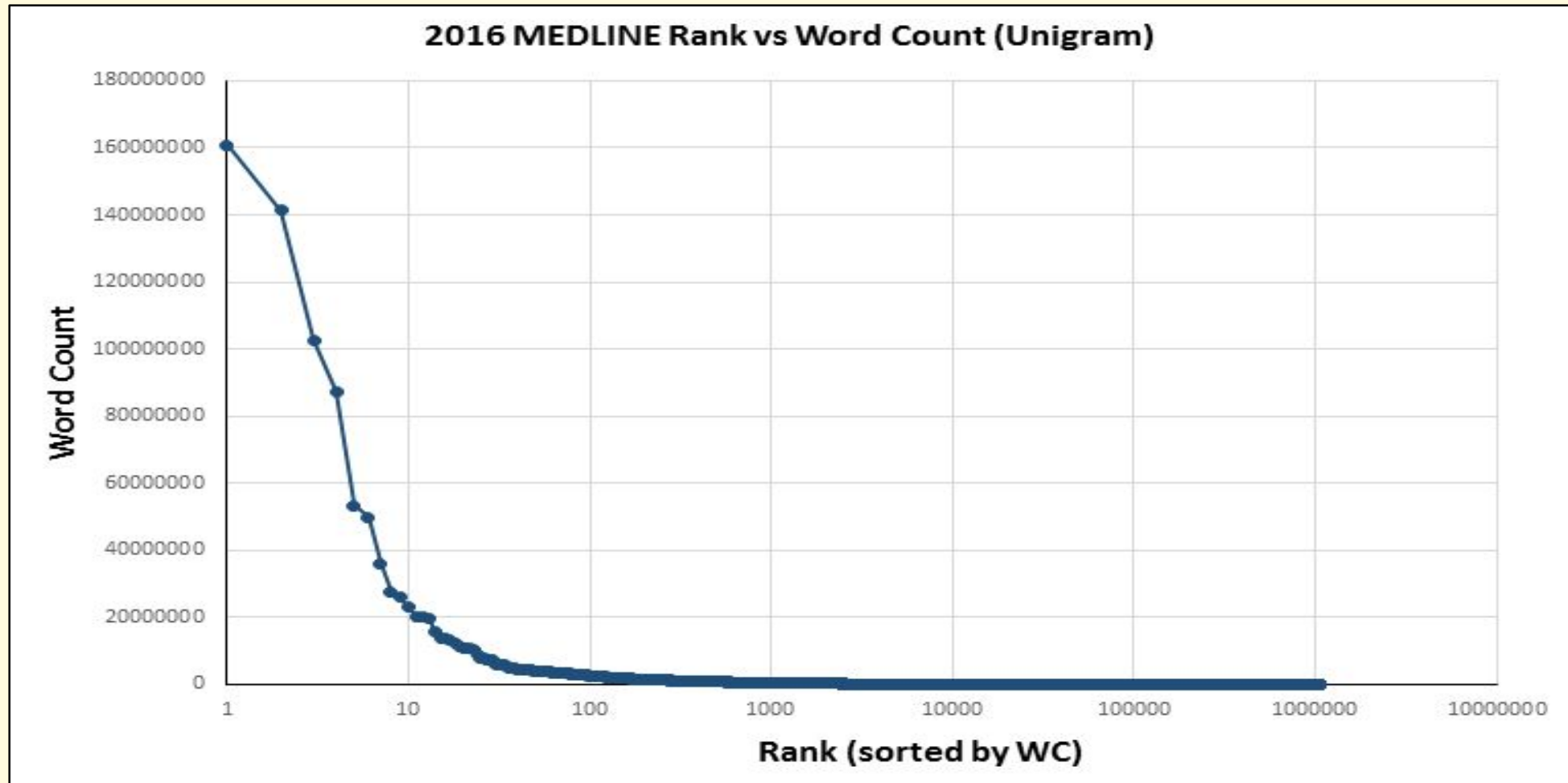
Frequency Spectrum of MEDLINE 2016



- The frequency spectrum of Alice in Wonderland, Word Frequency Distributions by R. Harald Baayen, 2001, Springer-Science + Business Media, B.V., P:10

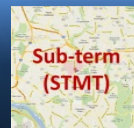
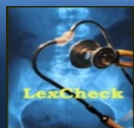
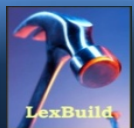
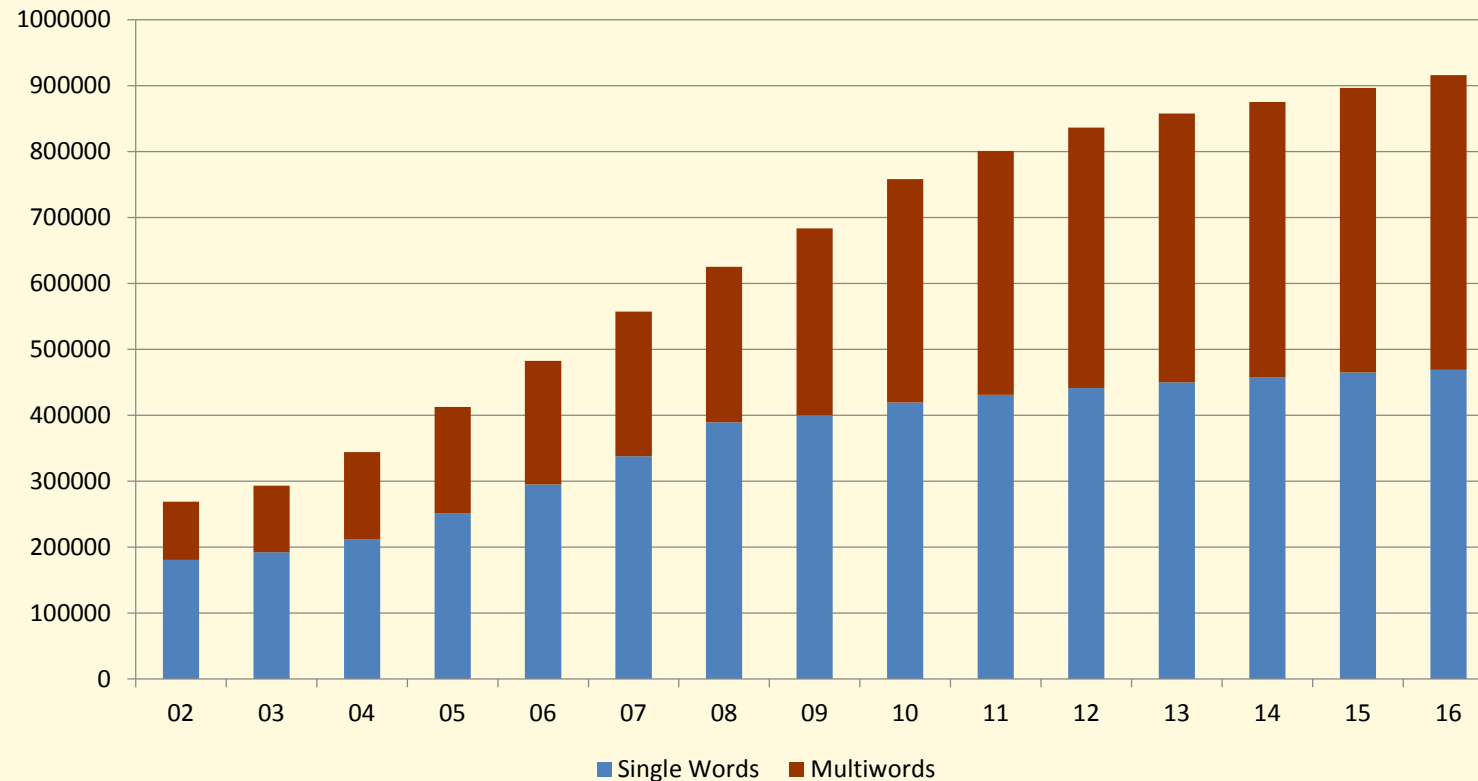


Word Frequency vs. Rank - MEDLINE 2016



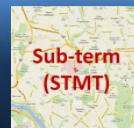
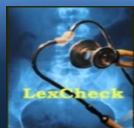
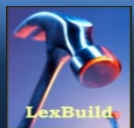
Lexicon Growth – 2002 to 2016

- 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%)



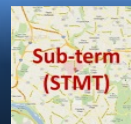
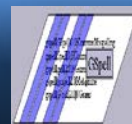
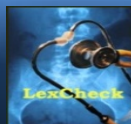
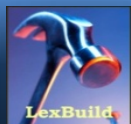
Future Lexicon Building

- 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

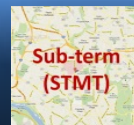
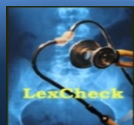
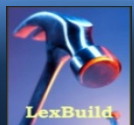


Multiwords Issues - Examples

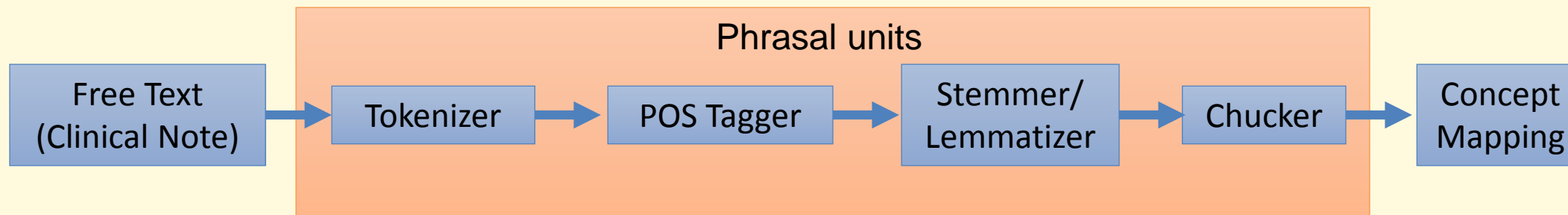
➤ Query Expansion

Synonym-key	Synonym-value	Query Expansion Example
...
perforated	perforation	perforated ear drum => perforation ear drum (Tympanic Membrane Perforation)
hot	warm	hot dog => warm dog
dog	canine	hot dog => hot canine
...

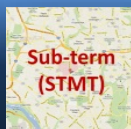
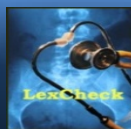
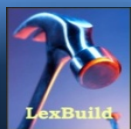
- The concept associated with a sentence often coincides with the longest multiword in the sentences (used in MetaMap)



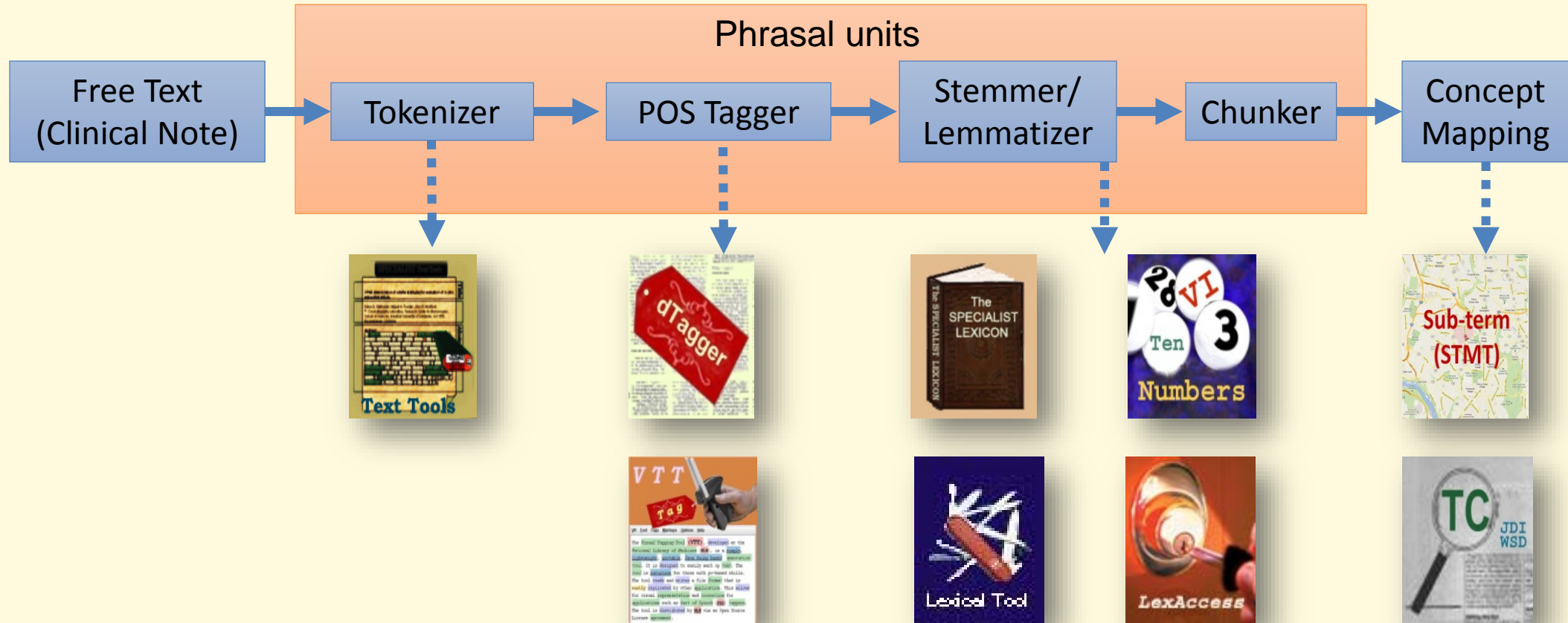
Multiwords in NLP



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



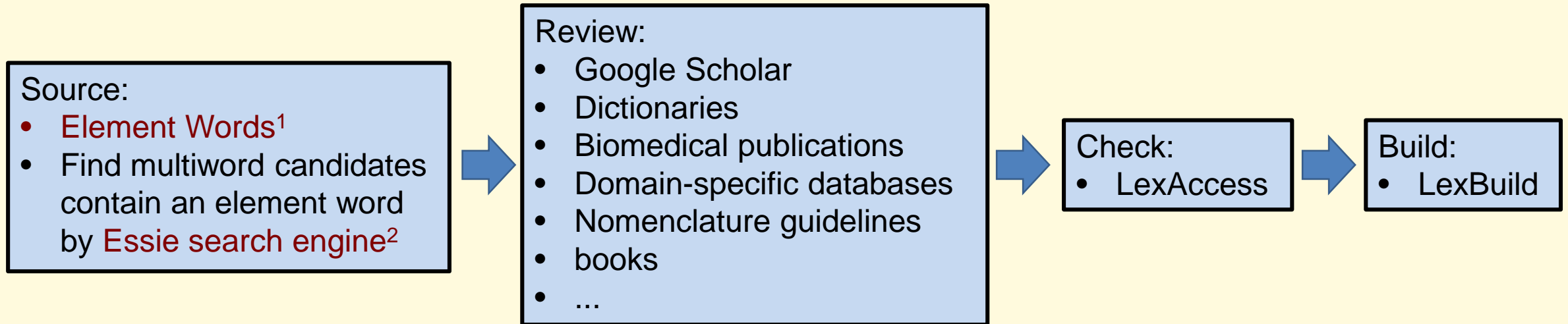
The SPECIALIST NLP Tools



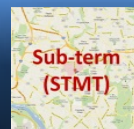
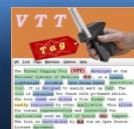
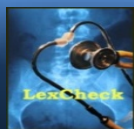
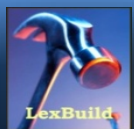
- Lexical Systems Group: <http://umlslex.nlm.nih.gov>
- The SPECIALIST NLP Tools: <http://specialist.nlm.nih.gov>



LexBuild Process (Computer-aided)

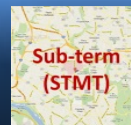
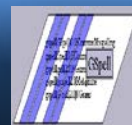
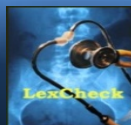
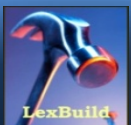


1. [“Using Element Words to Generate \(Multi\)Words for the SPECIALIST Lexicon”, Lu, Chris J.; Tormey, Destinee; McCreedy, Lynn; and Browne, Allen C. AMIA 2014 Annual Symposium, Washington, DC, November 15-19, 2014, p. 1499](#)
2. “Essie: A Concept-based Search Engine for Structured Biomedical Text”, N.C. Ide, R.F. Loane, D.D. Fushman, JAMIA, Vol. 14, Num. 3, May/June, 2007, p.253-263

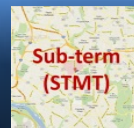
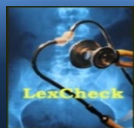
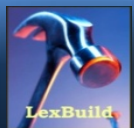
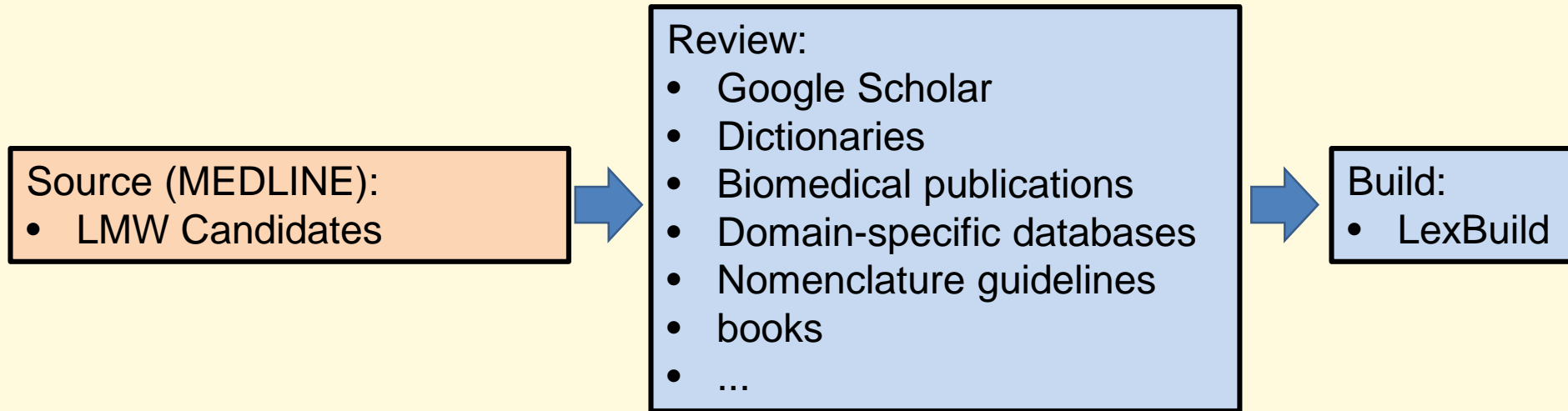


Issues of Element Word Approach

- 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

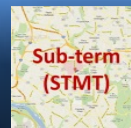
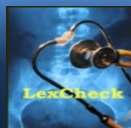
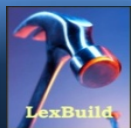


New LexBuild Process



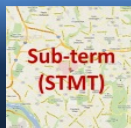
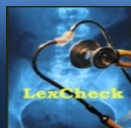
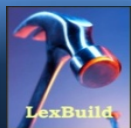
Project Objective

- A systematic way to add multiwords from MEDLINE to the SPECIALIST Lexicon:
 - Covers multiwords from the latest MEDLINE
 - Generates high precision multiword candidate list
 - To save time for linguists to build Lexicon



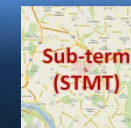
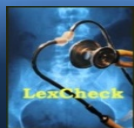
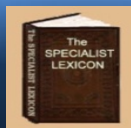
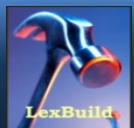
LexMultiword vs. Multiword Expression

- 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

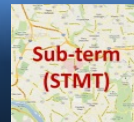
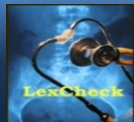
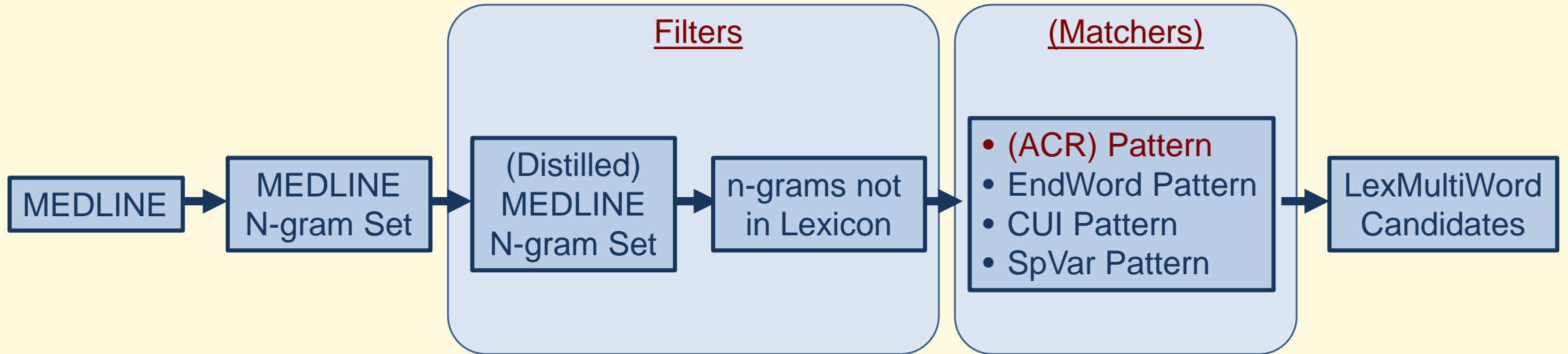


N-gram Approach

- 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.



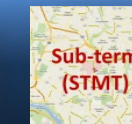
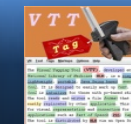
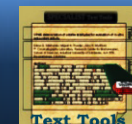
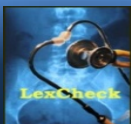
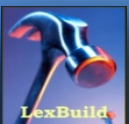
LMWs – Processes



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

N = 1	Unigram	to, be, or, not, to, be
N = 2	Bigram	to be, be or, or not, not to, to be
N = 3	Trigram	to be or, be or not, or not to, not to be



N-gram Requirements*

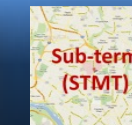
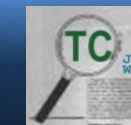
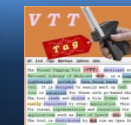
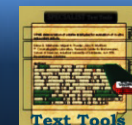
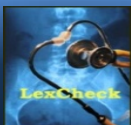
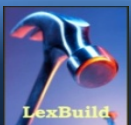
➤ Range of N:

- Lexicon.2016

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

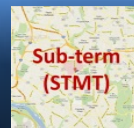
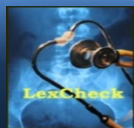
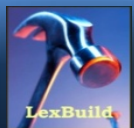
➤ Length: 50 (99.4562%) for Lexicon.2016

* “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

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

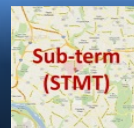
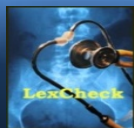
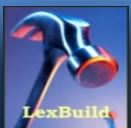


The MEDLINE N-gram Set

➤ Annual Public Releases:

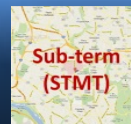
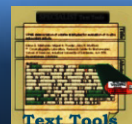
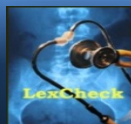
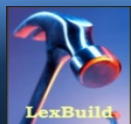
<http://umlslex.nlm.nih.gov/nGram>

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

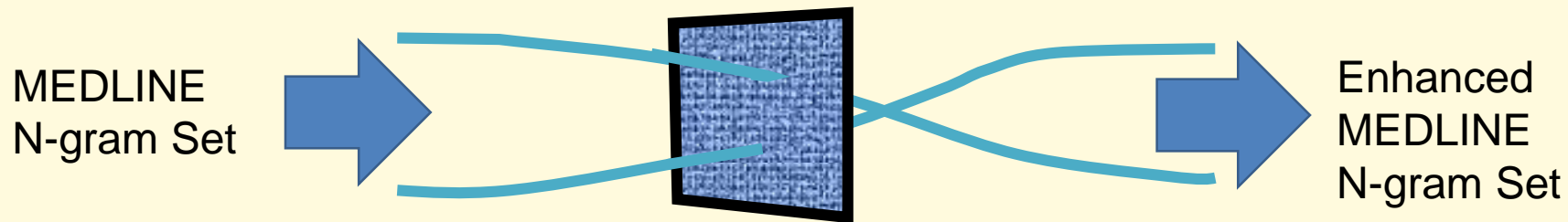


Enhanced N-gram Set?

- 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

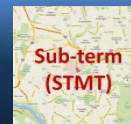
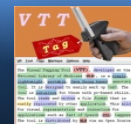
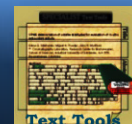
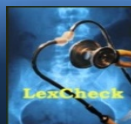
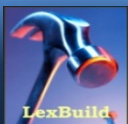


Filter

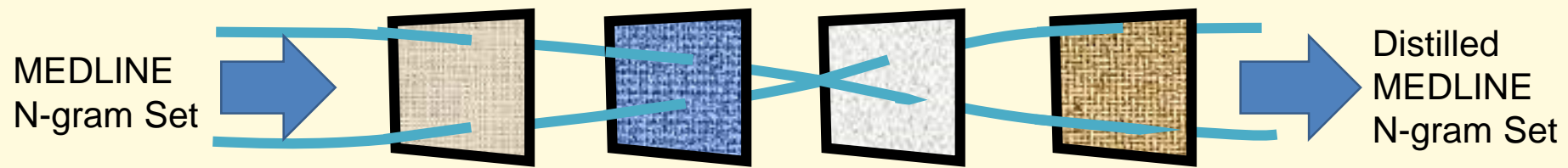


	Trap (not retrieved)	Pass (retrieved)
Valid (relevant)	FN	TP
Invalid (not relevant)	TN	FP

- 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

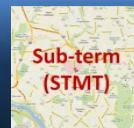
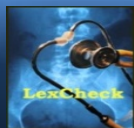
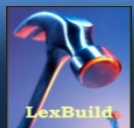


Serial Filters (High Accuracy)

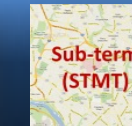
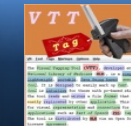
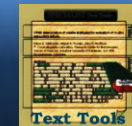
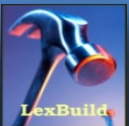
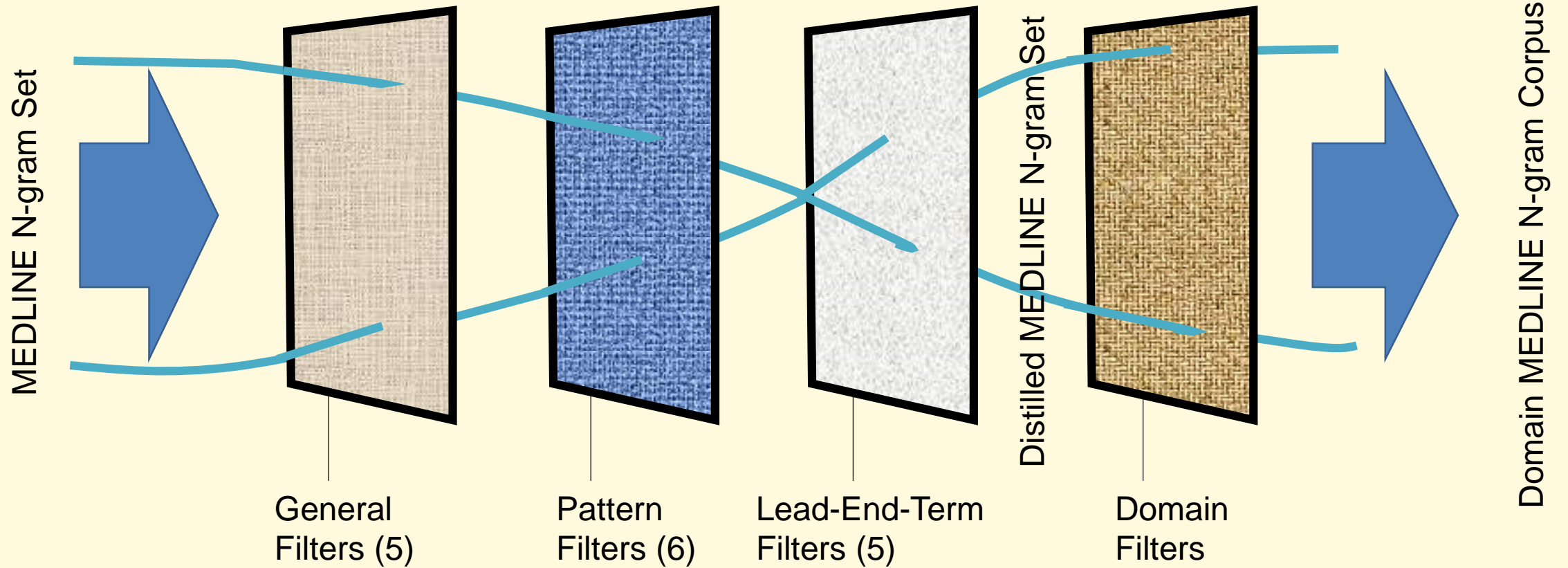


	N-gram	Filter-1	Filter-2	...	Filter-N	Distilled
Valid (TP)	V_0	V_1	V_2	...	V_n	V_n
Invalid (FP)	I_0	I_1	I_2	...	I_n	I_n

- A distilled n-gram set by filtering out invalid words.
- Applied high accuracy filter ($V_0 = V_1 = \dots = V_n$; $I_0 > I_1 > \dots > I_n$)
- Higher precision with same recall rate (if filter has high accuracy rate)
- N-gram Precision $n = V_n / (V_n + I_n)$
 - $= V_0 / (V_0 + I_n)$ V_n is same as V_0 (high accuracy)
 - $> V_0 / (V_0 + I_0)$ I_0 is bigger than I_n (high efficiency)
- N-gram Recall $n = V_n / (V_n + FN_n)$
 - $= V_n / (V_n + FN_0)$ FN_n is a constant (0), same as FN_0
 - $= V_0 / (V_0 + FN_0)$ V_n is same as V_0 (high accuracy)

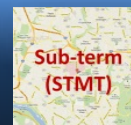
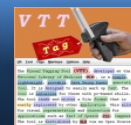
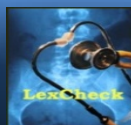


Distilled N-gram Set



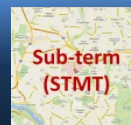
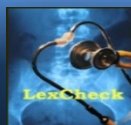
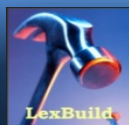
General Exclusive Filters

Filter	Accuracy On Lexicon (875,890)	Passing Rate N-gram Set	Accumulated Passing Rate	Trapped Examples
Pipe	100.0000% (0)	100.0000% (6)	100.0000%	<ul style="list-style-type: none"> • 38 44 (r • 33 37 Ag AgCl
Punctuation or space	100.0000% (0)	99.9977% (386)	99.9977%	<ul style="list-style-type: none"> • 1259147 3690494 = • 604567 2377864 +/-
Digit	99.9999% (1)	99.3141% (116,772)	99.3118%	<ul style="list-style-type: none"> • 1404799 2062240 2 • 239725 499064 95%
Number	99.9953% (41)	99.9760% (4,056)	99.2879%	<ul style="list-style-type: none"> • 2463066 3359594 two • 18246 20674 first and second
Digit and stopword	99.9993% (6)	99.1595% (142,067)	98.4534%	<ul style="list-style-type: none"> • 3155416 4125616 on the • 11180 12722 1, 2, and



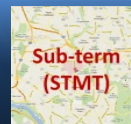
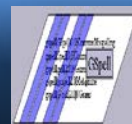
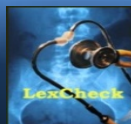
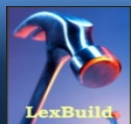
Pattern Exclusive Filters

Filter	Accuracy On Lexicon (875,890)	Passing Rate N-gram Set	Accumulated Passing Rate	Trapped Examples
Parenthetic acronym - (ACR)	100.0000% (0)	99.0232% (163,714)	97.4917%	<ul style="list-style-type: none"> • 33117 33381 chain reaction (PCR) • 30095 30315 polymerase chain reaction (PCR)
Indefinite article	99.9985% (13)	98.1703% (303,679)	95.7079%	<ul style="list-style-type: none"> • 270384 292590 a case • 40271 40512 A series
UPPERCASE colon	99.9999% (1)	99.4302% (92,841)	95.1625%	<ul style="list-style-type: none"> • 2069343 2070116 RESULTS: • 18015 18016 AIM: The
Disallowed punctuation	99.9978% (19)	99.3020% (113,073)	94.4983%	<ul style="list-style-type: none"> • 324405 719011 (n = • 86525 133350 (P < 0.05)
Measurement	99.9967% (29)	98.1947% (290,421)	92.7924%	<ul style="list-style-type: none"> • 154905 181001 two groups • 12160 15197 10 mg/kg
Incomplete	99.9999% (1)	97.8470% (340,109)	90.7945%	<ul style="list-style-type: none"> • 482021 1107869 (P • 25347 25992 years) with



Lead-End-Terms Exclusive Filters

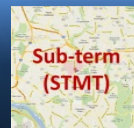
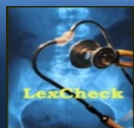
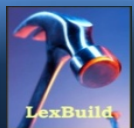
Filter	Accuracy On Lexicon (875,890)	Passing Rate N-gram set	Accumulated Passing Rate	Trapped Examples
Absolute Invalid Lead-Term	99.9947% (46)	73.0945% (4,158,702)	66.3658%	<ul style="list-style-type: none"> • 2780043 3451203 of a • 432921 434591 this study was
Absolute Invalid End-Term	99.9997% (3)	78.8984% (2,384,059)	52.3615%	<ul style="list-style-type: none"> • 1878109 3534031 patients with • 1062545 1261445 between the
Lead-End-Term	99.9992% (7)	99.9741% (2,312)	52.3480%	<ul style="list-style-type: none"> • 2578756 3106139 in a • 1733 1744 For one
Lead-Term no SpVar	99.9887% (99)	85.6678% (1,277,229)	44.8454%	<ul style="list-style-type: none"> • 658430 708246 to determine • 533913 554628 In addition,
End-Term no SpVar	99.9975% (22)	83.1945% (1,283,001)	37.3089%	<ul style="list-style-type: none"> • 1009451 1295670 number of • 726 734 (HPV) in



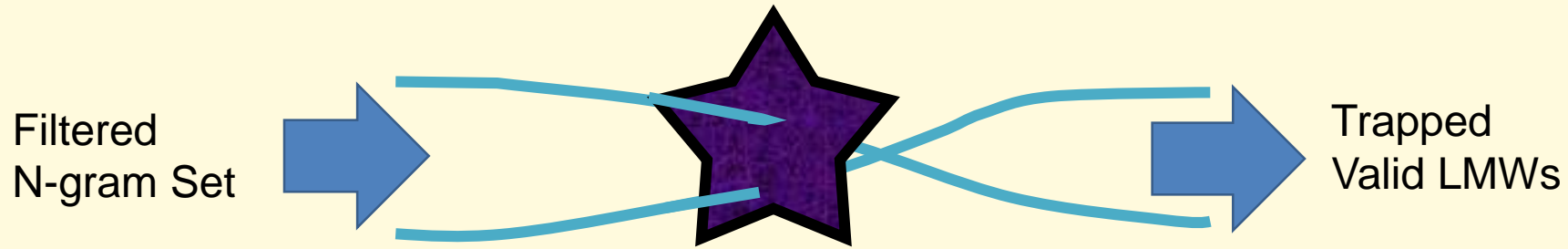
The Distilled MEDLINE N-gram Set

➤ Available to public: <http://umlslex.nlm.nih.gov/nGram>

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

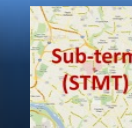
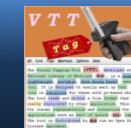
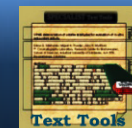
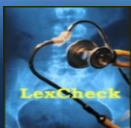
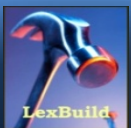


Matcher



	Trap (retrieved)	Pass (not retrieved)
Valid (relevant)	TP	FN
Invalid (not relevant)	FP	TN

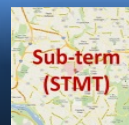
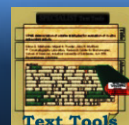
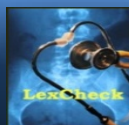
- **Parenthetic Acronym Pattern Matcher**
 - “computed tomography (CT)”, “magnetic resonance imaging (MRI)”, etc.
- **Spelling Variant Pattern Matcher**
 - Applied algorithm of SpVarNorm, Metaphone, edit distance, sorted distance, etc.
- **Metathesaurus CUI Pattern Matcher**
 - LMW candidate if a term 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.



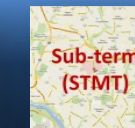
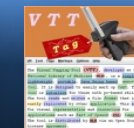
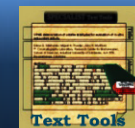
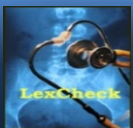
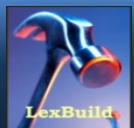
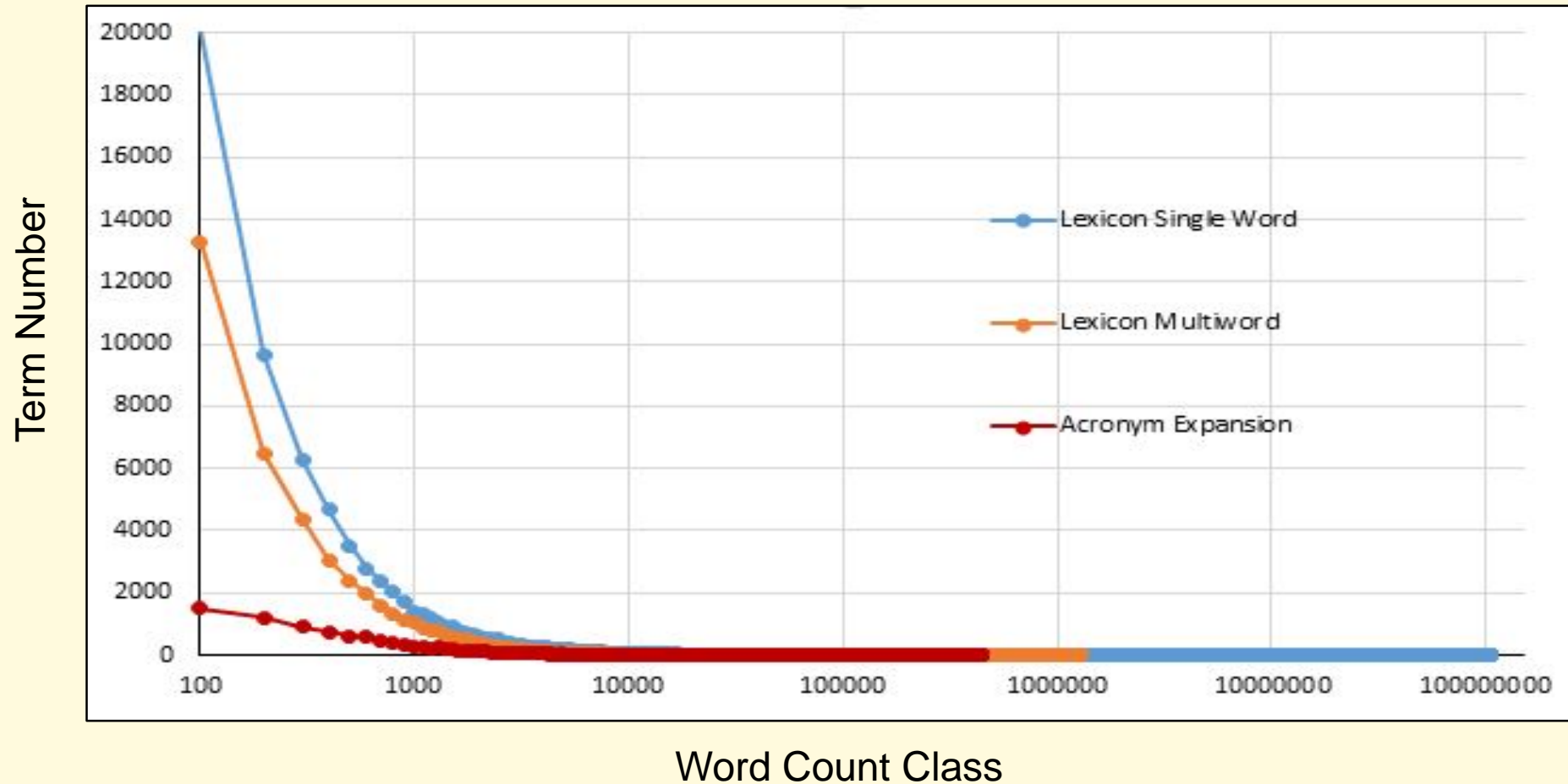
Practice Results

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

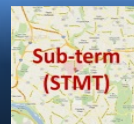
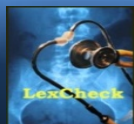
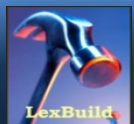
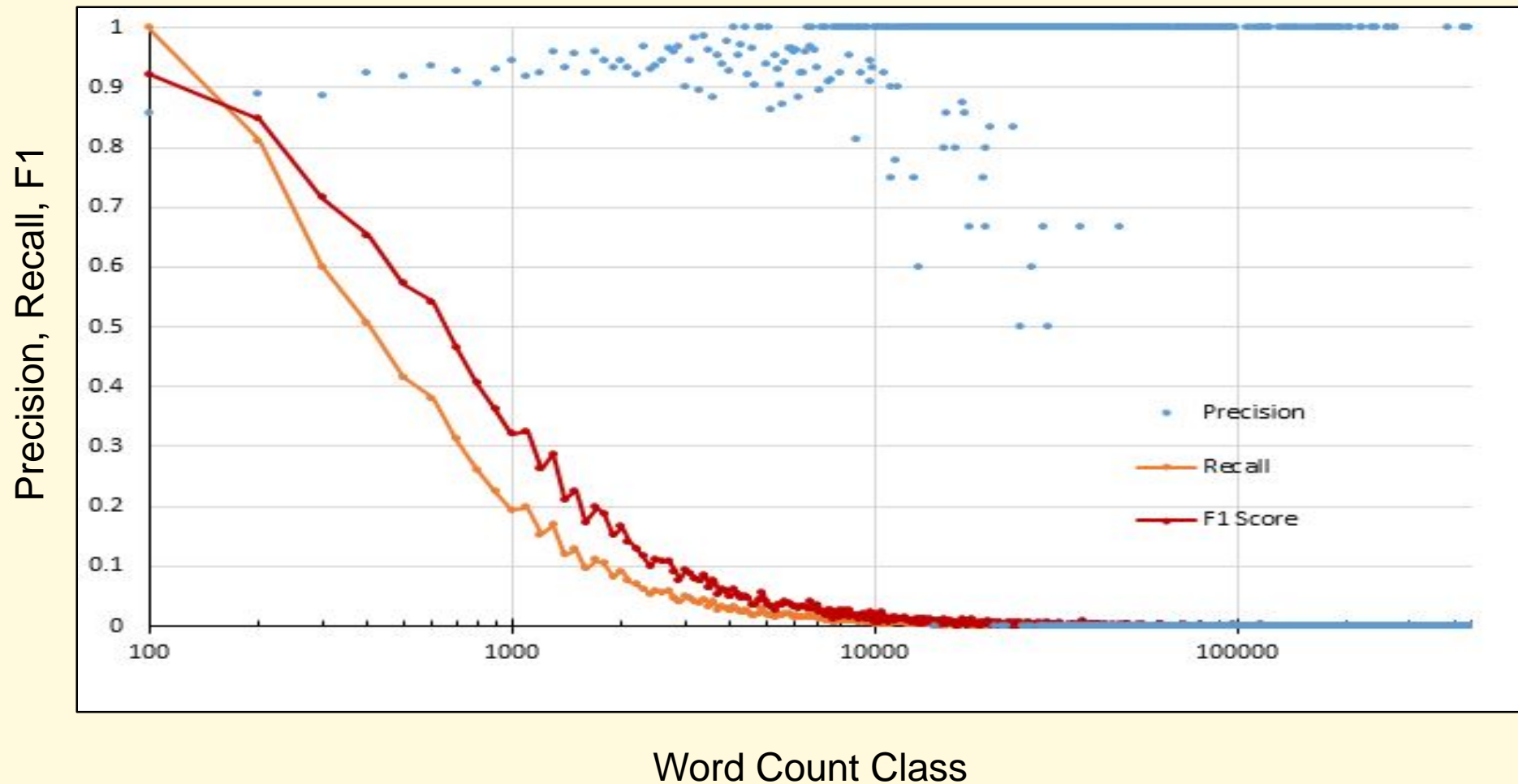
Case	Test Case - Model	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
1	Parenthetic Acronym - gold standard	14,805	1,870	0	0	0.8879	1.0000	0.9406	0.8879
2	Distilled MEDLINE N-gram Set (16 filters)	14,796	1,305	9	565	0.9189	0.9994	0.9575	0.9212
3	Spelling Variant Pattern matcher	7,509	482	7,296	1,388	0.9397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



Frequency Analysis – Valid Words Distribution



Frequency Analysis – PRF for AEP Model



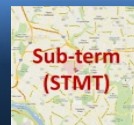
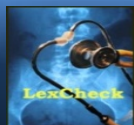
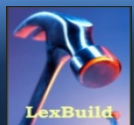
Frequency Analysis Summary

➤ 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.

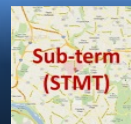
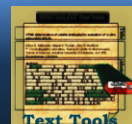
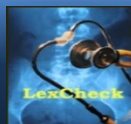
➤ Strategy

- 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



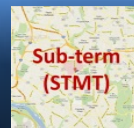
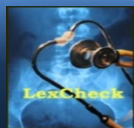
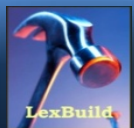
Summary

- 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



- Lexical Systems Group: <http://umlslex.nlm.nih.gov>
- The SPECIALIST NLP Tools: <http://specialist.nlm.nih.gov>

