The SPECIALIST NLP Tools
Generating Multiwords from MEDLINE

By: Dr. Chris J. Lu

The Lexical Systems Group
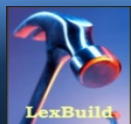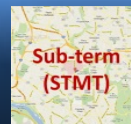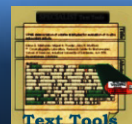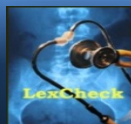
NLM – LHNCBC - CGSB

March, 2016

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

Table of Contents

➢ Introduction

➢ Objective and Approach

- Filters
- Matchers
- The Distilled MEDLINE N-gram Set

➢ Practice Results and Future Work

➢ Questions

Words in Lexicon

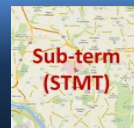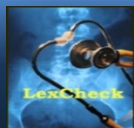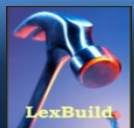➢ Part of speech, inflections, a special unit of meaning by itself
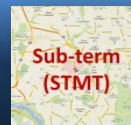
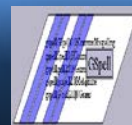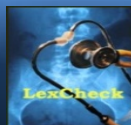- saw|noun|singular|E0054443

- saw|verb|infinitive|E0054444

- saw|verb|past|E0055007

Single Words vs. Multiwords

➢ Words include single words and multiwords
➢ Word boundary – space or tab
➢ Multiwords are words hat 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
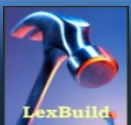
Lexicon Unigram Coverage - Unique Word

- ➢ Total unique word for MEDLINE (2015): 3,441,837
- ➢ Lexicon covers 12.17 % unigrams in MEDLINE

Types	Word Count	Percentage %	Accu. %
LEXICON (S)	294,765	8.5642%	8.5642%
NUMBER	62	0.0018%	8.5660%
DIGIT	80,925	2.3512%	10.9172%
MULTIWORD	43,155	1.2538%	12.1710%
NEW	3,022,930	87.8290%	100.0000%
Total	3,441,837		
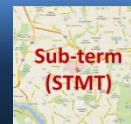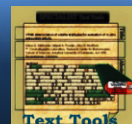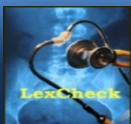
Lexicon Unigram Coverage – Word Count

➢ Total word count for MEDLINE (2015): 2,920,298,229
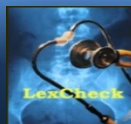➢ Lexicon covers > 99% unigrams from MEDLINE

Types	Word Count	Percentage %	Accu. %
LEXICON	2,728,200,576	93.4220%	93.4220%
NUMBER	8,283,899	0.2837%	93.7057%
DIGIT	136,010,712	4.6574%	98.3631%
MULTIWORD	19,618,128	0.6718%	99.0349%
NEW	28,184,914	0.9651%	100.0000%
Total	2,920,298,229		
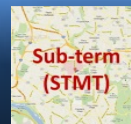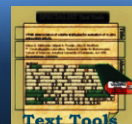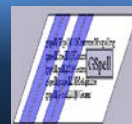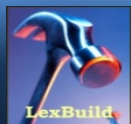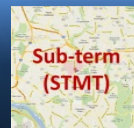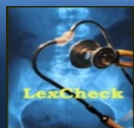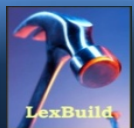
Word Frequency vs. Rank - MEDLINE 2015
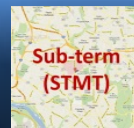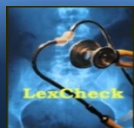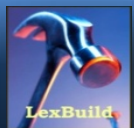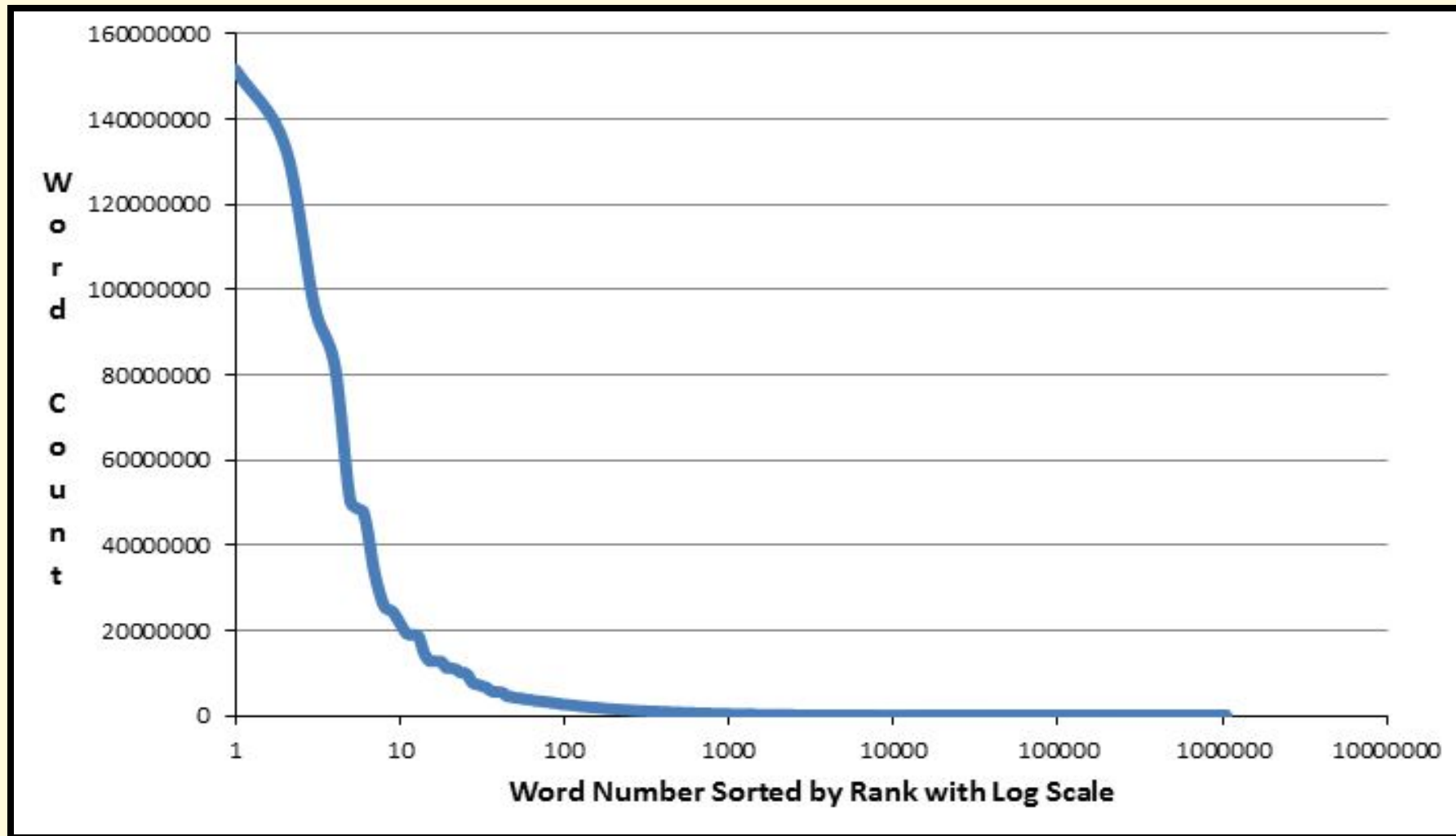
Lexicon.2016

- ➢ 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

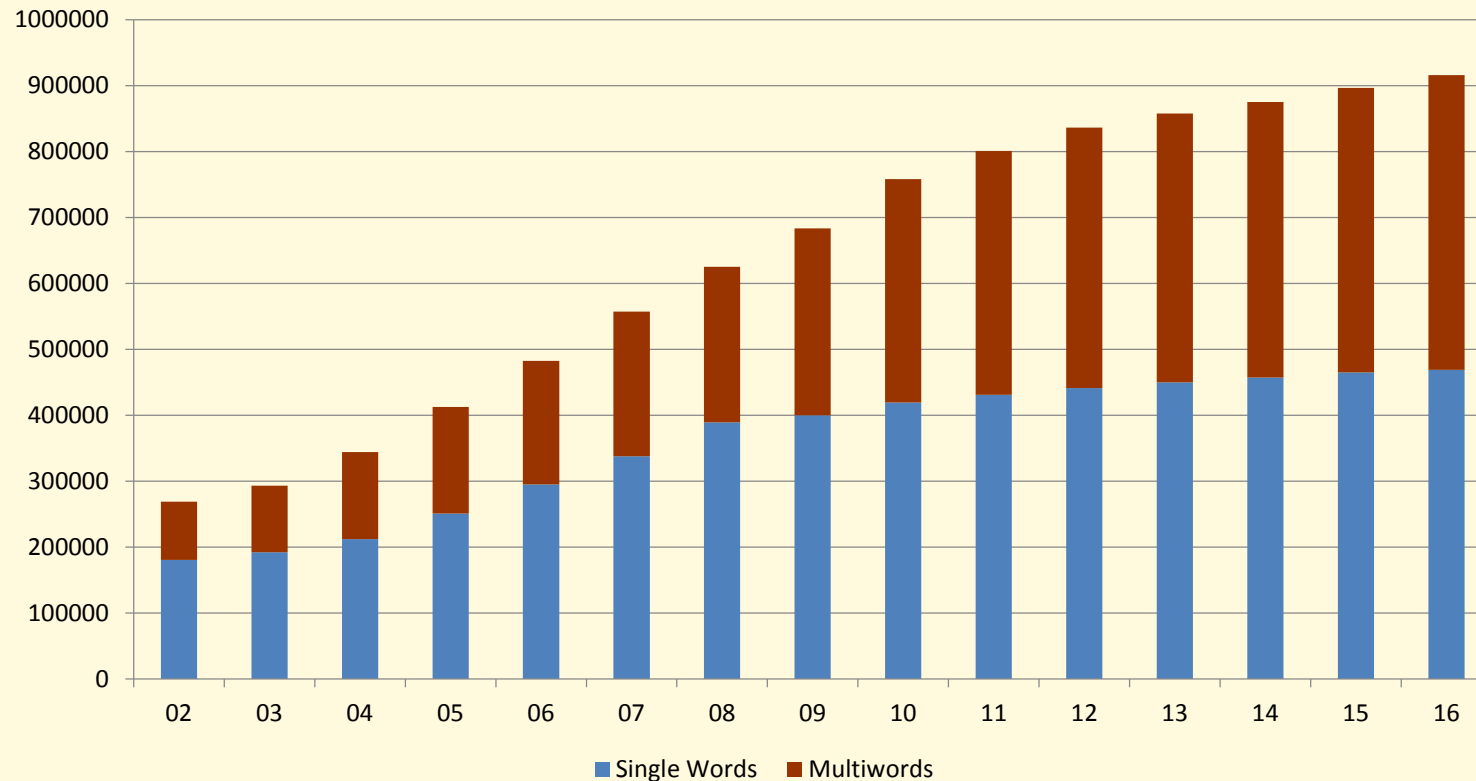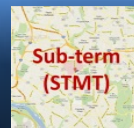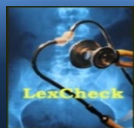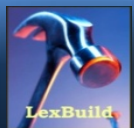➢ Lexicon single words: high coverage
➢ Lexicon multiwords (LMWs): increasing growth
➢ Multiwords acquisition is the key for future Lexicon building

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	warn	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)
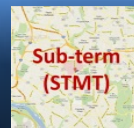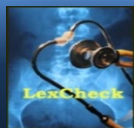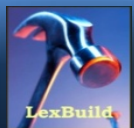
Multiwords in NLP
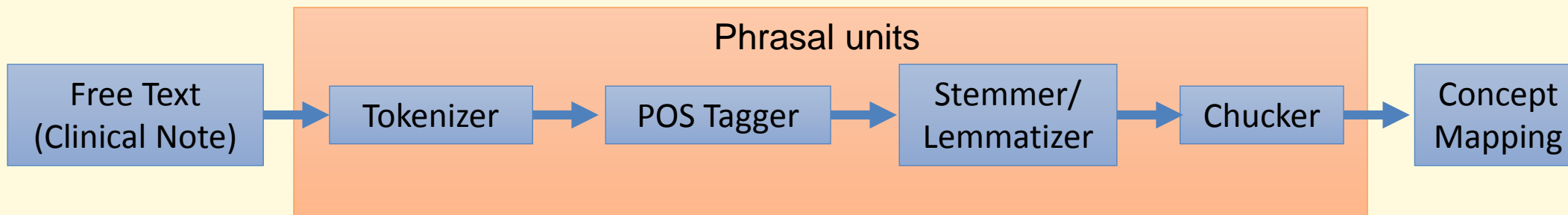
- ➢ Identify multiwords as phrasal units directly
- ➢ Reduce part-of-speech ambiguity
- ➢ Improve stemming and lemmatization
- ➢ Better concept mapping results

The SPECIALIST NLP Tools

LexBuild Process (Computer-aided)

Source:
- Element Words[1]
- Find multiword candidates contain an element word by Essie search engine[2]

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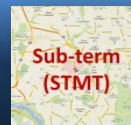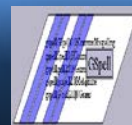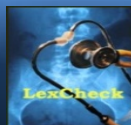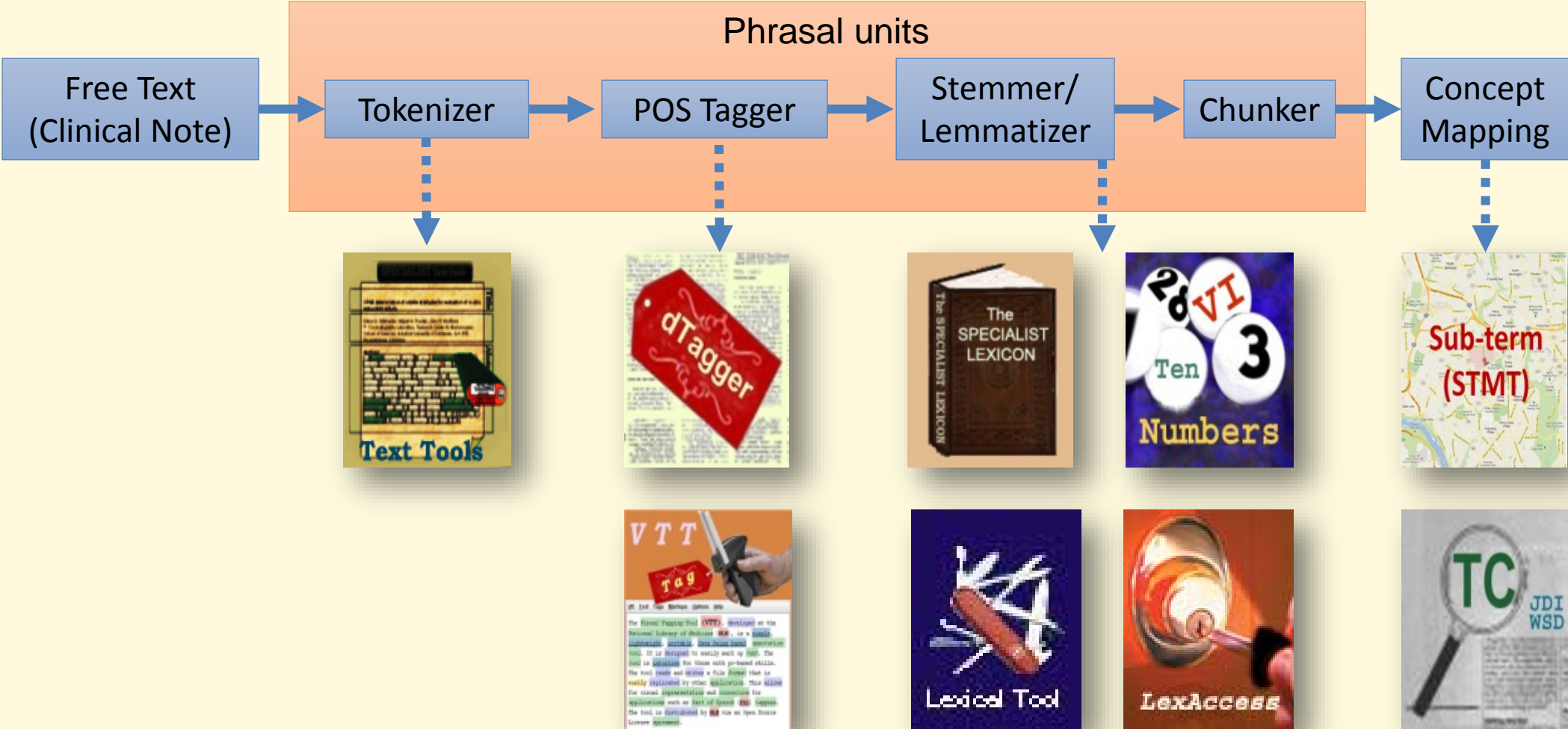- ...

Check:
- LexAccess

Build:
- LexBuild
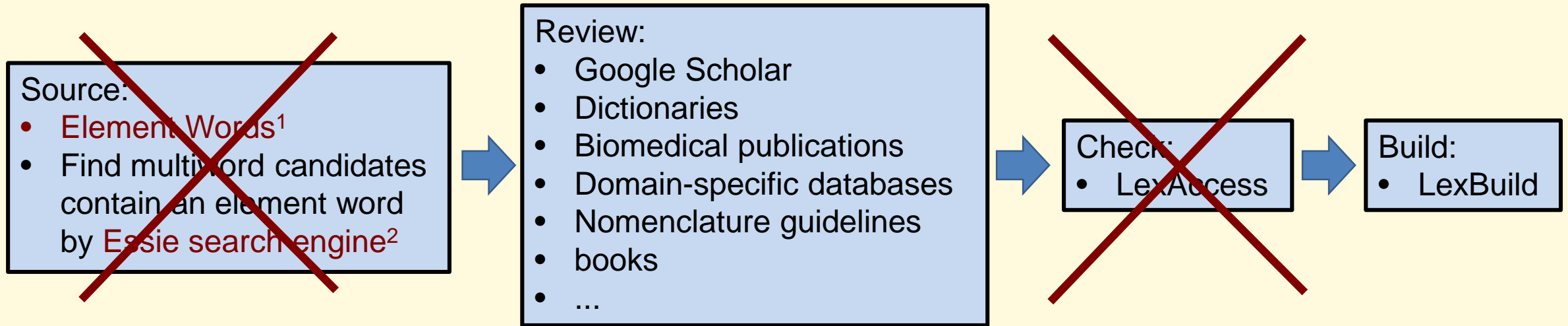
1. "Using Element Words to Generate (Multi)Words for the SPECIALIST Lexicon",
   Lu, Chris J.; Tormey, Destinee; McCreedy, Lynn; and Browne, Allen C.
   AMIA 2014 Annual Symposium, Washington, DC, November 15-19, 2014, p. 1499
2. "Essie: A Concept-based Search Engine for Structured Biomedical Text",
   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

Source (MEDLINE):
- LMW Candidates

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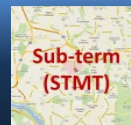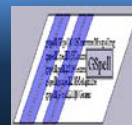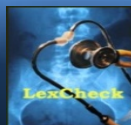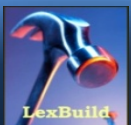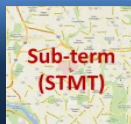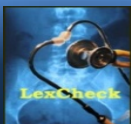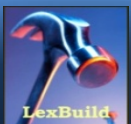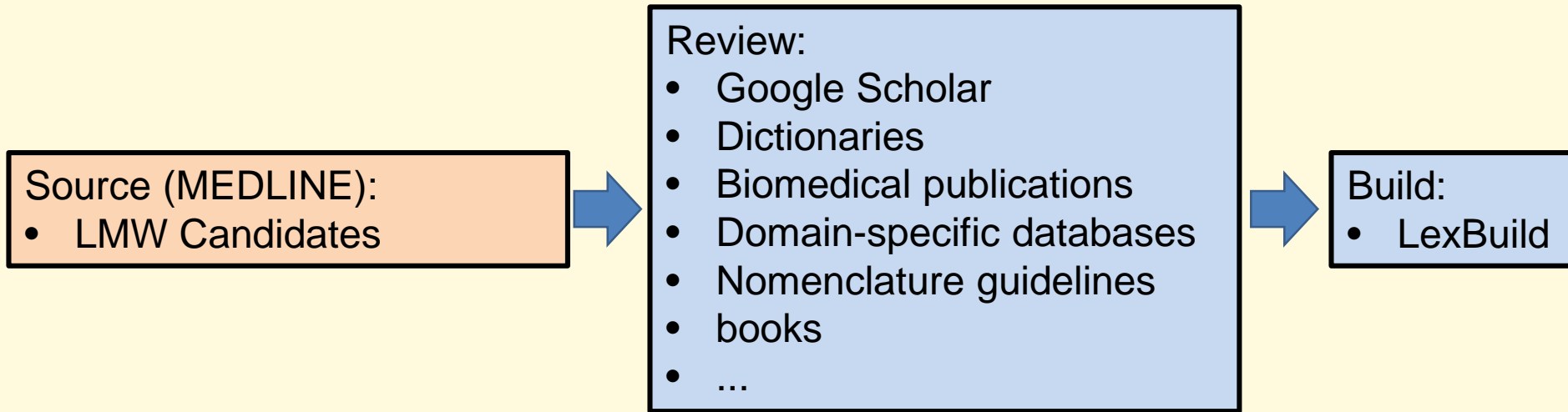
Build:
- LexBuild

Project Objective

➢ A systematic way to add multiwords form MEDLINE to the SPECIALIST Lexicon:

- Covers high frequency multiwords from the latest MEDLINE
- Generates high precision multiword candidate list
  - To save time for linguist to build Lexicon
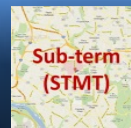
LexMultiword vs. Multiword Expression

➢ 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.
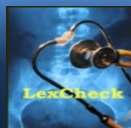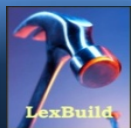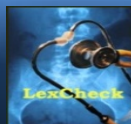➢ Non-decomposable idioms (beyond the score of the Lexicon)
  • "kick the bucket", "shoot the breeze", etc.
➢ Design goal is set to five-grams to reach coverage above 99%
  • Most MWE research only focus on bi-grams or tri-grams
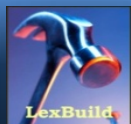
N-gram Approach

- ➢ 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

MEDLINE → MEDLINE N-gram Set →

Filters

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

(Matchers)

- (ACR) Pattern
- EndWord Pattern
- CUI Pattern
- SpVar Pattern
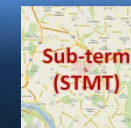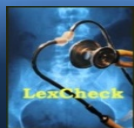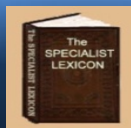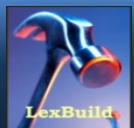
→ LexMultiWord Candidates
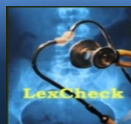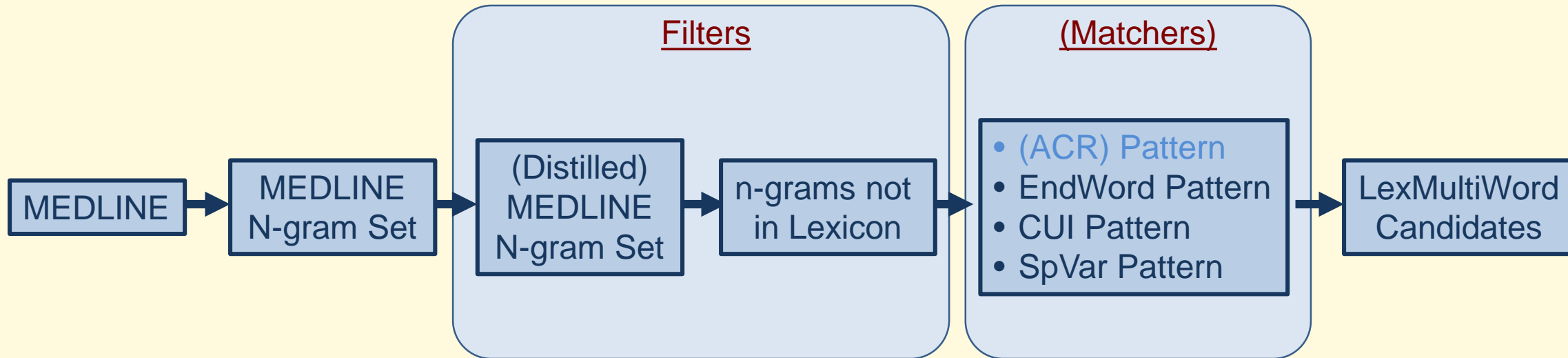
N-gram Requirements

➢ 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
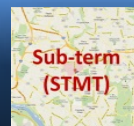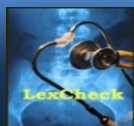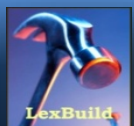
The MEDLINE N-gram Set - Specifications

N-grams	2014	2015
MEDLINE files	1-746	1-779
Max. length	50	50
Min. WC	30	30
Min. DC	1	1
Total documents	22,356,869	23,343,329
Total sentences	126,612,705	134,834,507
Total tokens	2,610,209,406	2,786,085,158
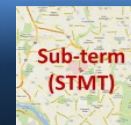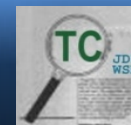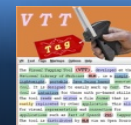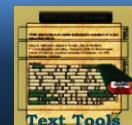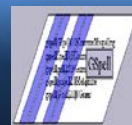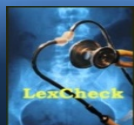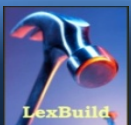
The MEDLINE N-gram Set

➤ Annual Public Releases:
  http://umlslex.nlm.nih.gov/nGram

N-grams	2014	2015
unigrams	804,382	843,206
bigrams	4,587,349	4,845,965
trigrams	6,287,536	6,702,194
four-grams	3,799,377	4,082,612
five-grams	1,545,175	1,674,715
n-gram set	17,023,819	18,148,692
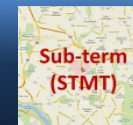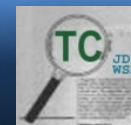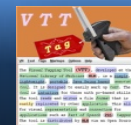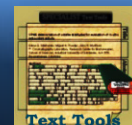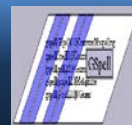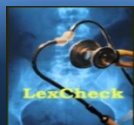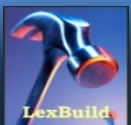
Enhanced N-gram Set?

- ➢ 17 ~ 18 million is still a big number
- ➢ Reduce the size by filtering out invalid multiwords:
  - increase precision
  - without sacrificing recall
  - distilled MEDLINE n-gram set
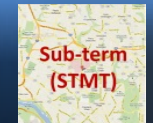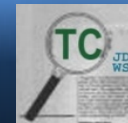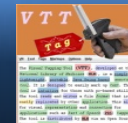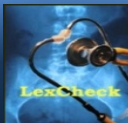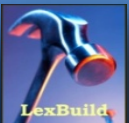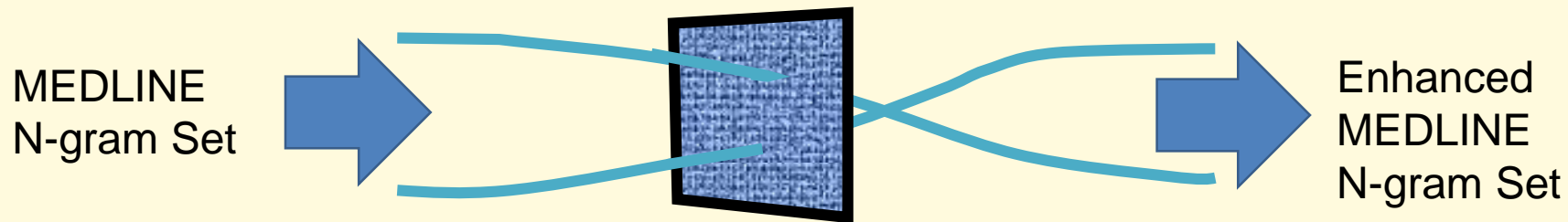
Filter

	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
    - = pass rate
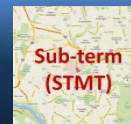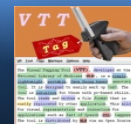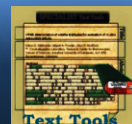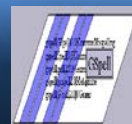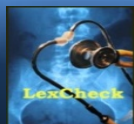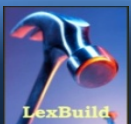
Serial Filters (High Accuracy)

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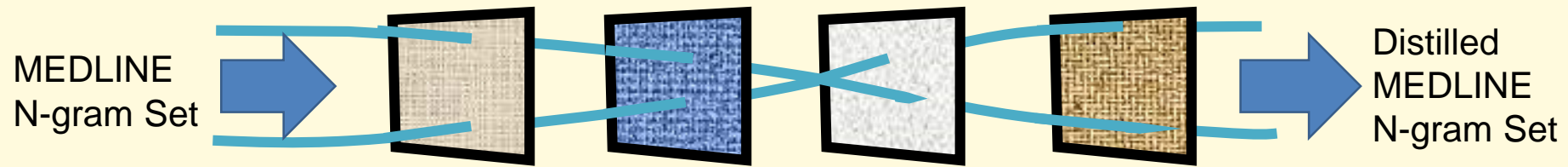
	N-gram	Filter-1	Filter-2	...	Filter-N	Distilled
Valid (TP)	V_0	V_1	V_2	...	V_n	V_n
Invalid (FP)	I_0	I_1	I_2	...	I_n	I_n

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

Distilled N-gram Set

MEDLINE N-gram Set

General Filters (5)

Pattern Filters (6)

Lead-End-Term Filters (5)

Distilled MEDLINE N-gram Set

Domain Filters

Domain MEDLINE N-gram Corpus
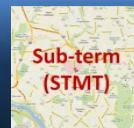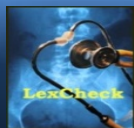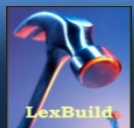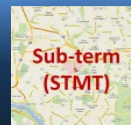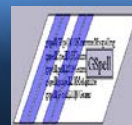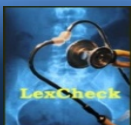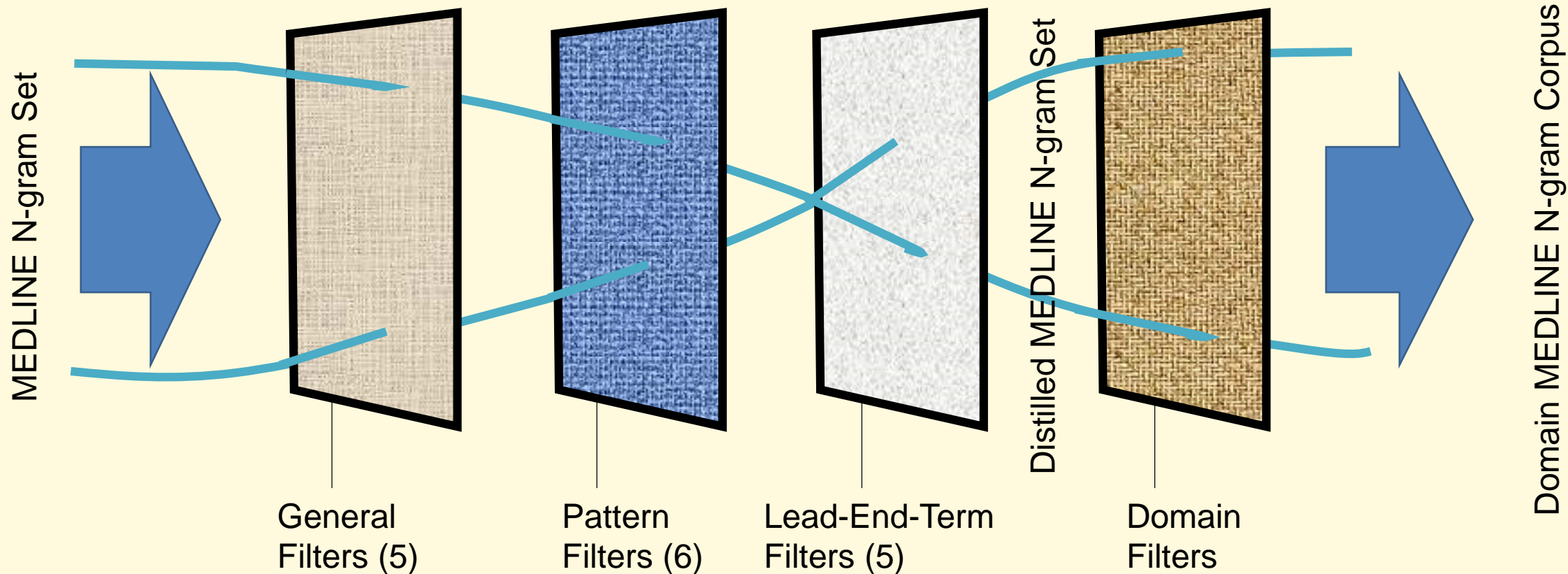
General Exclusive Filters

Filter	Accuracy (875,890)	Pass Rate N-gram Set	Accumulated Pass Rate	Trapped Examples
Pipe	100.0000% (0)	100.0000% (6)	100.0000%	• 38|44|(|r| • 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
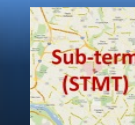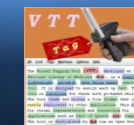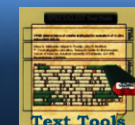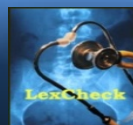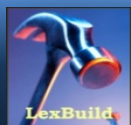
Pattern Exclusive Filters

Filter	Accuracy (875,890)	Pass Rate N-gram Set	Accumulated Pass Rate	Trapped Examples
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

Lead-End-Terms Exclusive Filters

| Filter	Accuracy (875,890)	Pass Rate N-gram set	Accumulated Pass Rate	Trapped Examples
Absolute Invalid Lead-Term	99.9947% (46)	73.0945% (4,158,702)	66.3658%	• 2780043|3451203|of a • 432921|434591|this study was
Absolute Invalid End-Term	99.9997% (3)	78.8984% (2,384,059)	52.3615%	• 1878109|3534031|patients with • 1062545|1261445|between the
Lead-End-Term	99.9992% (7)	99.9741% (2,312)	52.3480%	• 2578756|3106139|in a • 1733|1744|For one
Lead-Term no SpVar	**99.9887% (99)**	85.6678% (1,277,229)	44.8454%	• 658430|708246|to determine • 533913|554628|In addition,
End-Term no SpVar	99.9975% (22)	83.1945% (1,283,001)	**37.3089%**	• 1009451|1295670|number of • 726|734|(HPV) in
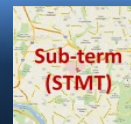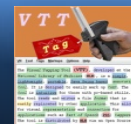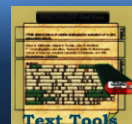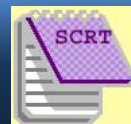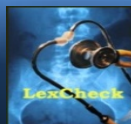
The Distilled MEDLINE N-gram Set

➢ Available to public:
http://umlslex.nlm.nih.gov/nGram

N-grams	2014	2015
unigrams	804,382	843,206
bigrams	4,587,349	4,845,965
trigrams	6,287,536	6,702,194
fourthgrams	3,799,377	4,082,612
fifthgrams	1,545,175	1,674,715
N-gram Set	17,023,819	18,148,692
Distilled N-gram Set	6,351,392	6,793,561

(~ 37%)

Matcher

Filtered N-gram Set → Trapped Valid LMWs
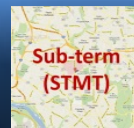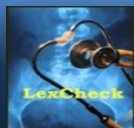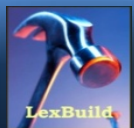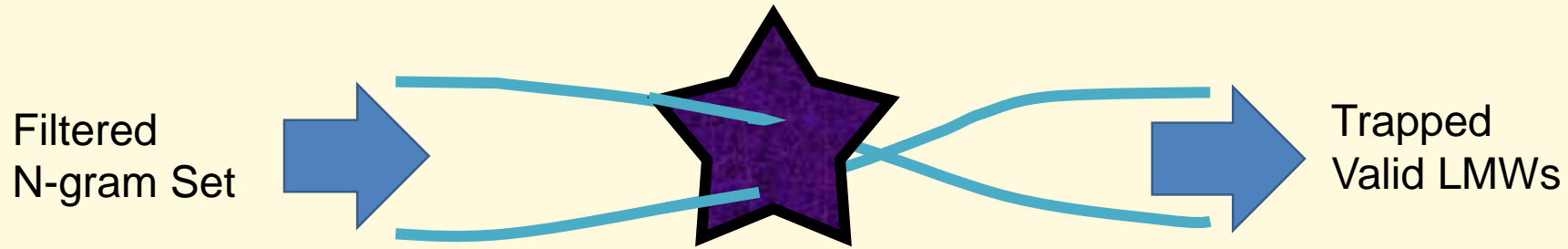
	Trap (retrieved)	Pass (not retrieved)
Valid (relevant)	TP	FN
Invalid (not relevant)	FP	TN

➤ **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.
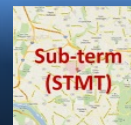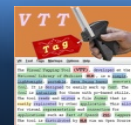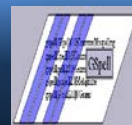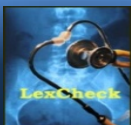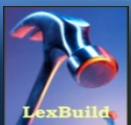
Practice Results

➢ Baseline: 14,400 LMW Candidates from (ACR) matcher, tagged by linguists

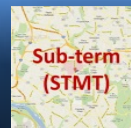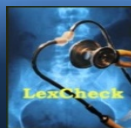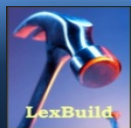| Case	Test Case - Model	TP	FP	Precision	Recall	F1
1	**Parenthetic Acronym - gold standard**	13,170	1,230	0.9146	1.0000	0.9554
2	**Distilled MEDLINE N-gram Set (16 filters)**	13,165	795	0.9431	0.9996	0.9705
3	**Spelling Variant Pattern matcher**	6,609	283	0.9589	0.5018	0.6589
4	**Metathesaurus CUI Pattern matcher**	8,678	512	0.9443	0.6589	0.7762
5	**EndWord Pattern matcher**	1,587	108	0.9363	0.1205	0.2135
6	**SpVar + CUI + Distilled**	4,993	127	0.9752	0.3791	0.5460
7	**SpVar + CUI + EndWord + Distilled**	690	5	0.9928	0.0524	0.0995
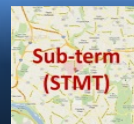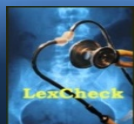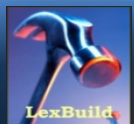
Summary

- All filters have accuracy rate above 99.99% (tested on Lexicon)
- Obtain the distilled MEDLINE n-gram set at passing rate of ~37.4%
  - smaller data set
  - better precision
  - similar recall
  - $\Rightarrow$ The recall rate between the Lexicon test set (0.999671) and baseline (0.999620) 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
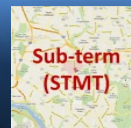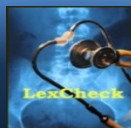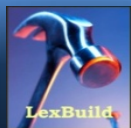
Future Work

➢ Continuously enhance filters and matchers for LexBuilding on multiwords

➢ More practice by applying 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