The SPECIALIST NLP Tools Multiwords
The Distilled MEDLINE N-gram Set

By: Dr. Chris J. Lu

The Lexical Systems Group

NLM – LHNCBC - CGSB

Oct., 2015

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

Table of Contents

Natural Language Processing (NLP)

- Map terms to concepts (meaning)
- Challenges: many to many mapping:

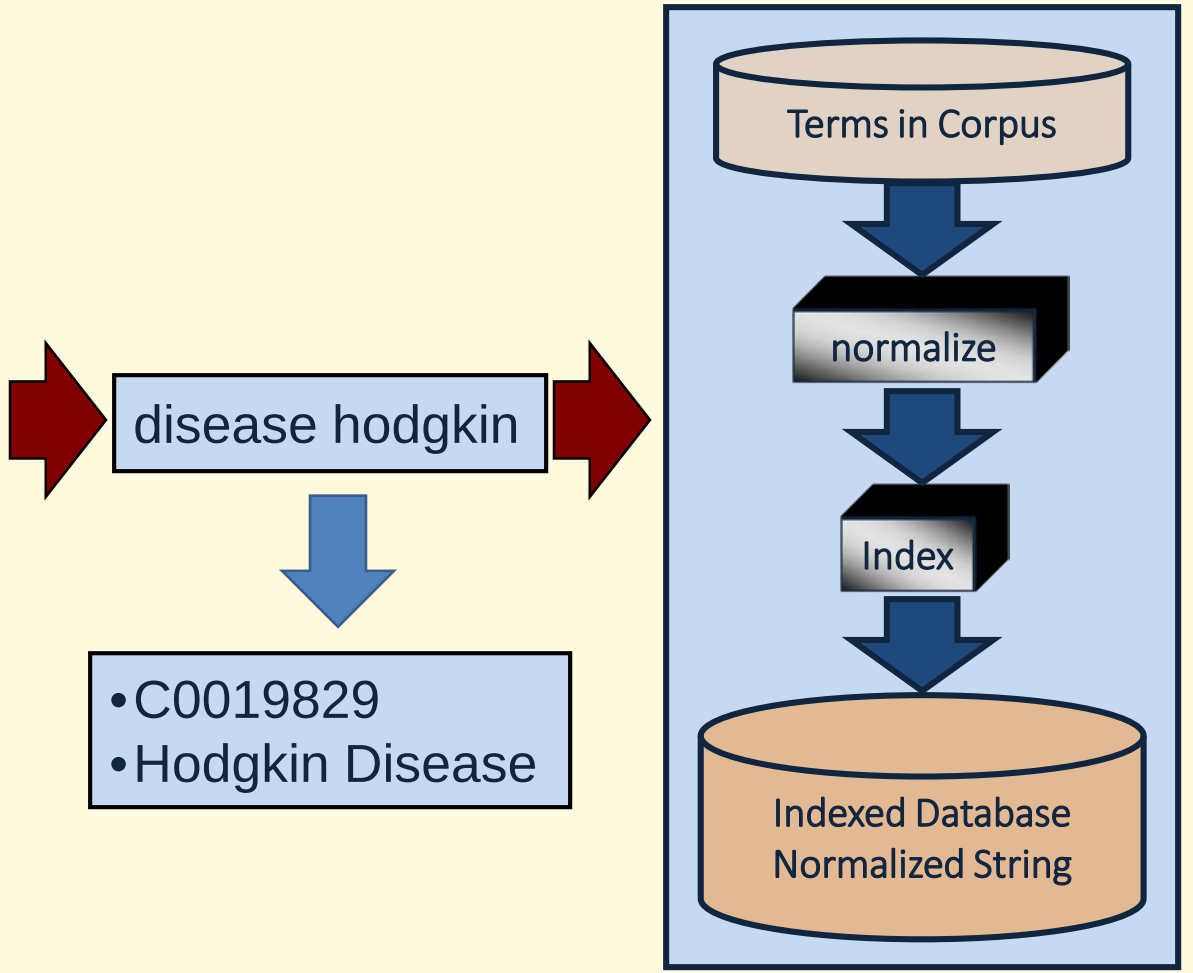
Terms	Concepts
• cold	• Cold Temperature|C0009264 • Common Cold|C0009443 • Cold Therapy|C0010412 • Cold Sensation|C0234192 • …
• cold • Cold Temperature • Cold Temperatures • Cold (Temperature) • Temperatures, Cold • Low temperature • low temperatures • …	• Cold Temperature|C0009264

NLP – Concept Mapping

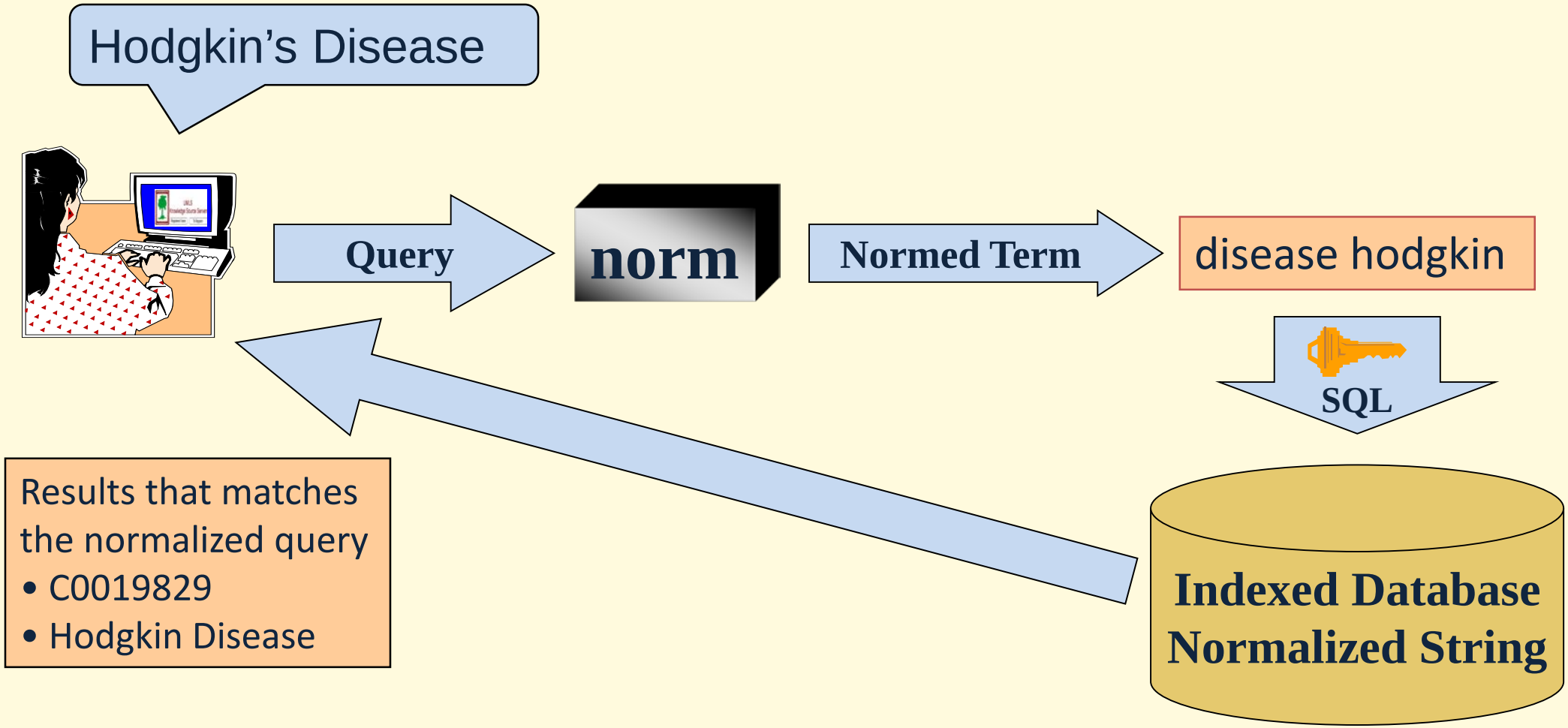
- Normalization:
  - A term might have a great deal of lexical variations, such as inflectional variants, spelling variants, synonyms, abbreviations (expansions), cases, ASCII conversion, etc.
  - Normalize different forms of a concept to a same form
- Query Expansion:
  - Expand a term to its equal terms, such as subterm substitution of synonyms, derivational variants, spelling variants, abbreviations, etc.
  - To increase recall
- POS tagger:
  - Assign part of speech to a single word or multiword in a text
  - To increase precision
- Others…

NLP - Norm

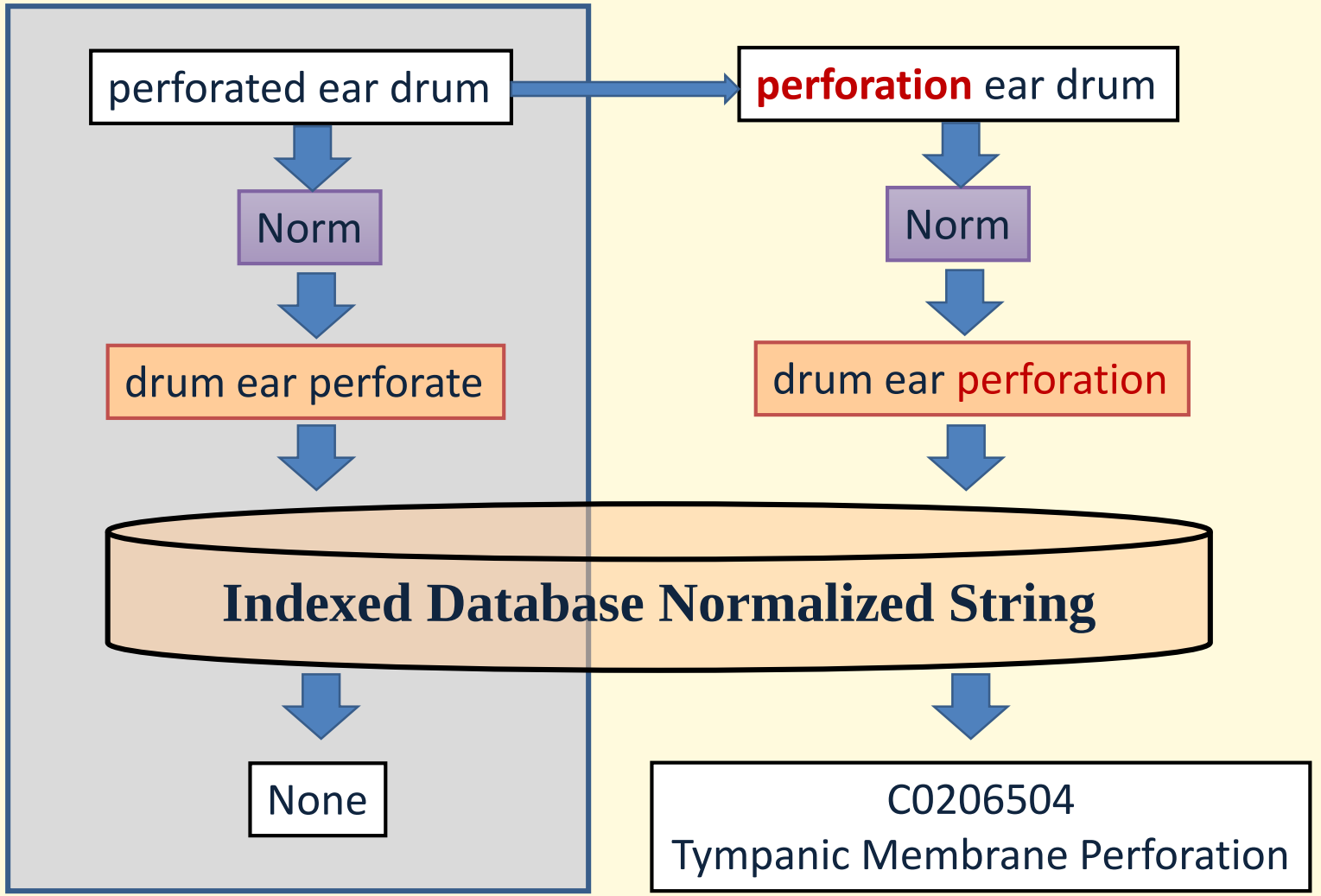
- Hodgkin Disease
- HODGKINS DISEASE
- Hodgkin's Disease
- Disease, Hodgkin's
- HODGKIN'S DISEASE
- Hodgkin's disease
- Hodgkins Disease
- Hodgkin's disease NOS
- Hodgkin's disease, NOS
- Disease, Hodgkins
- Diseases, Hodgkins
- Hodgkins Diseases
- Hodgkins disease
- hodgkin's disease
- Disease;Hodgkins
- Disease, Hodgkin
- …



NLP – Norm (Cont.)



NLP – Query Expansion

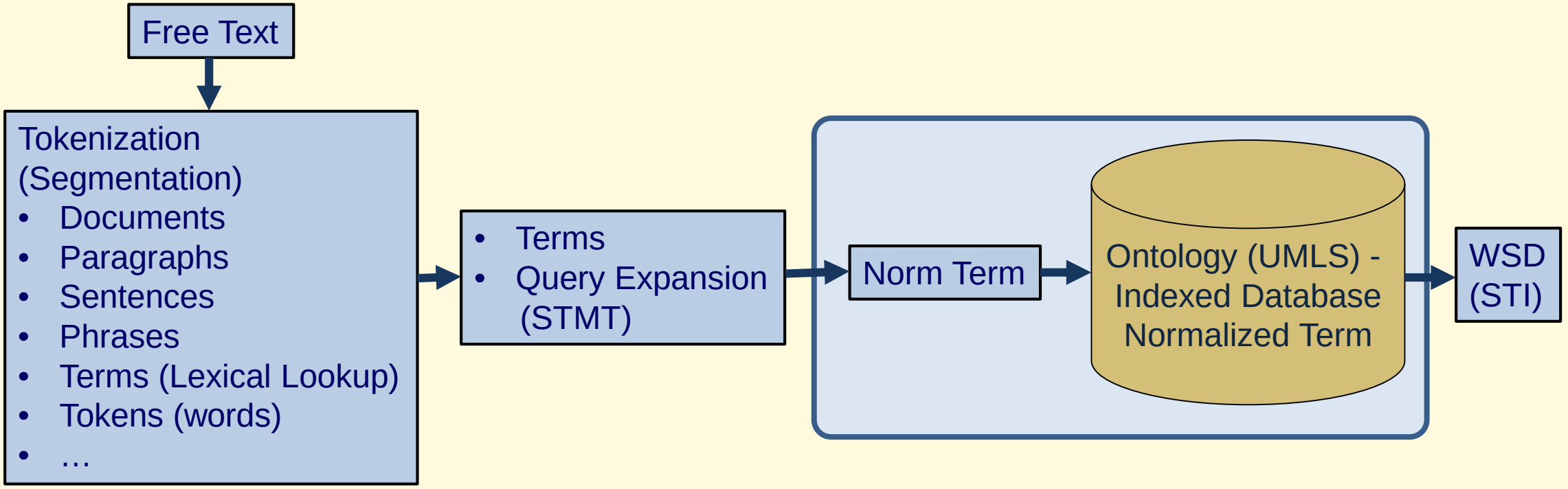


LVG - Lexical Variants Generation

- To increase recall & precision

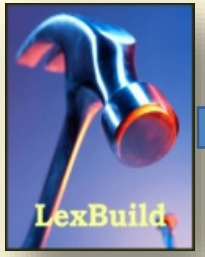
	Recall - Query Expansion	Precision - POS Tagging
Inputs	perforated ear drum	saw
UMLS-CUI	None	• C1947903|verb|see • C0183089|noun|saw (device)
Process	perforation ear drum	noun
UMLS-CUI	C0206504	• C0183089
Preferred Term	Tympanic Membrane Perforation	saw (device)

NLP – Concept Mapping Model



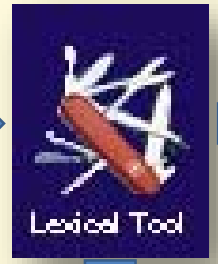
The SPECIALIST NLP Tools

LexBuild



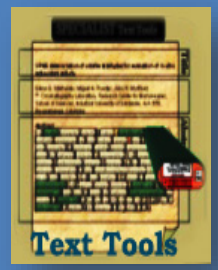
The SPECIALIST LEXICON

Lexical Tools



LVG

Text Tools



NLP Applications

The SPECIALIST NLP Tools by LSG

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

Lexicon Coverage – by Word Count

- Total word count for MEDLINE (2014): 2,725,710,505
- Lexicon covers > 98% from MEDLINE

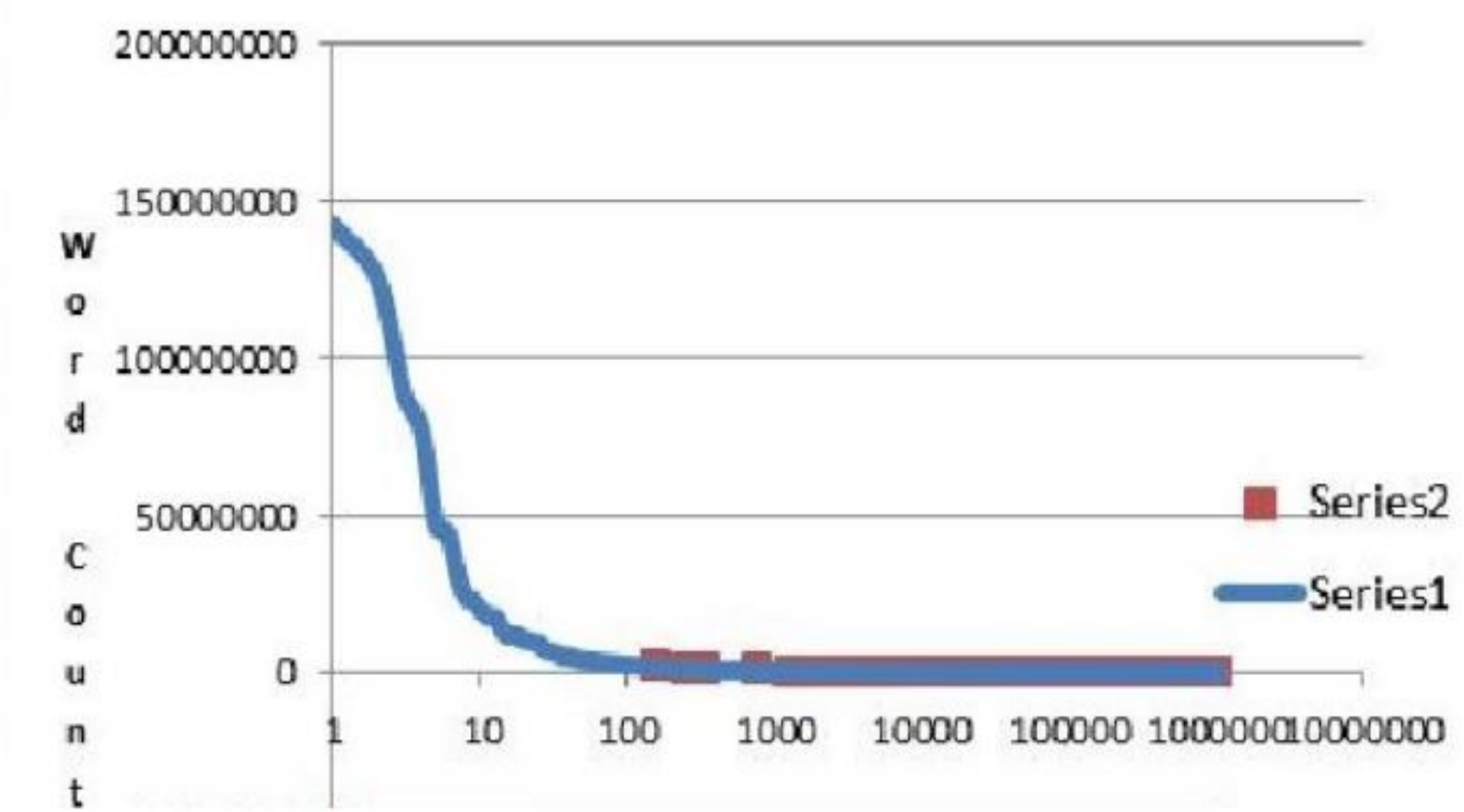
Types	Word Count	Percentage %	Accu. %
LEXICON	2,542,758,048	93.2879%	93.2879%
NUMBER	7,797,019	0.2861%	93.5740%
DIGIT	126,635,190	4.6460%	98.2200%
MULTIWORD	18,549,715	0.6805%	98.9005%
NEW	29,970,533	1.0995%	100.0000%
Total	2,725,710,505		

Lexicon Coverage - by Unique Word

- Total unique word for MEDLINE (2014): 3,264,205
- Lexicon covers 11 ~ 12 % words in MEDLINE
- The rest of ~87.5% is the long tail (multiwords)

Types	Word Count	Percentage %	Accu. %
LEXICON	291,271	8.9232%	8.9232%
NUMBER	61	0.0019%	8.9251%
DIGIT	75,406	2.3101%	11.2352%
MULTIWORD	42,045	1.2881%	12.5233%
NEW	28,55,422	87.4768%	100.0000%
Total	3,264,205		

Frequency Spectrum of MEDLINE 2014



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

Words in Lexicon

- Part of speech, inflection, lexical meaning
  - saw|noun|singular|E0054443

  - saw|verb|infinitive|E0054444

  - saw|verb|past|E0055007

LexMultiwords (LMW)

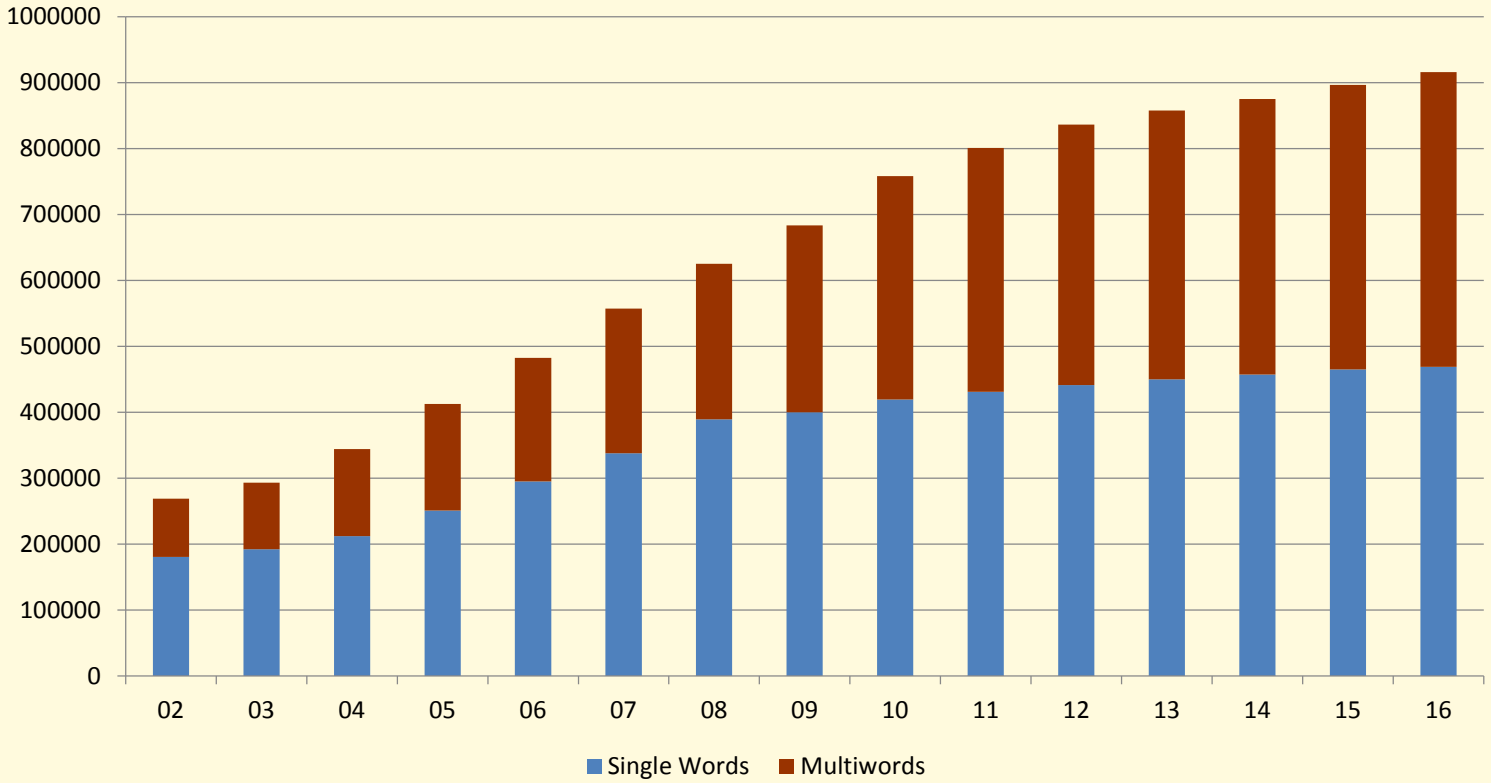
- A multiword is a word contains space(s)

- Multiwords are used extensively in biomedical domain
- Multiwords are an essential ingredient and play a key role for the success of NLP task
- Precise recognition of word boundaries and identify multiwords benefit disambiguation and improves the accuracy in information extraction
  - Example:

Synonym-key	Synonym-value	Query Expansion Example
…	…	…
perforated	perforation	perforated ear drum => **perforation** ear drum
hot	warn	hot dog => **warm** dog
dog	canine	hot dog => hot **canine**
…	…	…

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%)



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

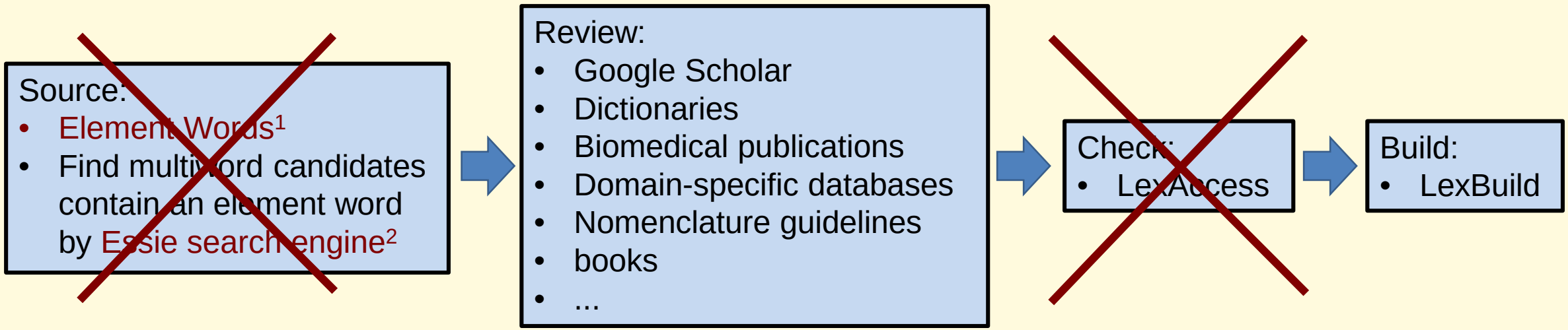
Build:
- LexBuild

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

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

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

N-gram Model 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.

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

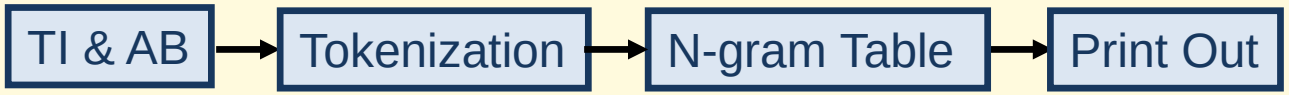
- Range of N:
  - Lexicon.2014

N	WC	Accumulated WC
1	457,335 (52.2615%)	457,335 (52.2615%)
2	281857 (32.2089%)	739,192 (84.4704%)
3	93011 (10.6287%)	832,203 (95.0991%)
4	29905 (3.4174%)	862,108 (98.5165%)
5	8358 (0.9551%)	870,466 (99.4716%)
6	2846 (0.3252%)	873,312 (99.7968%)
…	…	…

- Length: 50 (> 99.5508%) for Lexicon.2014
- Others: frequency (WC and DC)

MEDLINE N-gram Set Generation

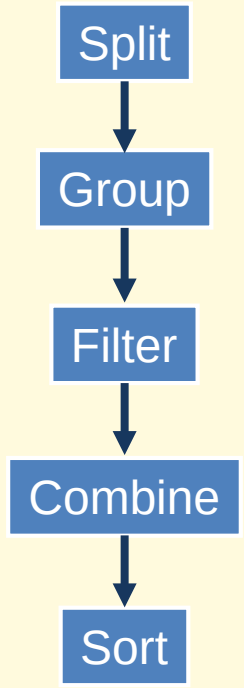
- 2014 Release:
  - Collect titles and abstracts from 22,356,869 MEDLINE documents
  - Tokenize titles and abstracts to 126,612,705 sentences
  - Parse sentences into n-grams
  - Print out: DC|WC|N-gram



  - Issues:
    - Real data: sentences tokenizer (unrecognized pattern < 0.01%)
    - Big data: many issues to generate n-grams when n >= 3 (key > 10^9, HashMap key: 2^{30} -1)

MEDLINE N-gram Set

- Approach – Split, Group, Filter, Combine, and Sort[1]
- Example - fourthgrams (automatic):
  - Split MEDLINE documents into 12 sections and get the fourthgrams for each section
  - Group fourthgrams from all 12 sections with specified (10) alphabetic range, such as a-c, c-e, e-f, etc.
  - Apply WC (> 30) filter on all 10 groups
  - Combine all 10 alphabetic ranges groups to n-gram set
  - Sort



1. "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, 20145

SGFCS Example: 4-Grams

PMID- 961033
PMID- 961032
PMID- 961031
TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, …

Split: 4-grams

MEDLINE Section 1/N

DC	WC	4-gram
…	…	…
11424	11843	as a result of
19407	21425	in the absence of
19592	20514	on the basis of
…	…	…

MEDLINE Section 2/N

DC	WC	4-gram
…	…	…
10784	11171	as a result of
20700	23036	in the absence of
18238	19184	on the basis of
…	…	…

. . .

MEDLINE Section N/N

. . .

Group (Alphabetically)

- **a-c: 182,666,123**

DC	WC	4-gram
…	…	…
106784	111389	as a result of
164023	168018	an important role in
…	…	…

- **f-k: 170,063,871**

DC	WC	4-gram
…	…	…
165302	182850	in the absence of
170501	192895	for the treatment of
…	…	…

- **k-p: 72,660,681**

DC	WC	4-gram
…	…	…
154616	163362	on the basis of
40956	42858	no effect on the
…	…	…

Filter-Combine-Sort

DC	WC	4-gram
…	…	…
170501	192895	for the treatment of
165302	182850	in the absence of
164023	168018	an important role in
154616	163362	on the basis of
106784	111389	as a result of
40956	42858	no effect on the
…	…	…

Exclusive Filter – WC (>= 30)

- 2014 MEDLINE N-gram Set:

N-grams	N	No. of N-grams	No. of n-grams (WC >= 30)	Pass Rate
unigrams	1	21,530,469	804,382	3.74%
bigrams	2	205,868,398	4,587,349	2.23%
trigrams	3	703,148,136	6,287,536	0.89%
four-grams	4	1,295,096,308	3,799,377	0.29%
five-grams	5	1,665,248,566	1,545,175	0.09%
n-gram set	1-5	3,890,891,877	17,023,819	0.44%

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

Frequency on N-grams

- Low Frequency:
  - Typos
  - Address, name, etc.
  - Measurement, range, etc.

- High Frequency co-occur words?

  a sequence of words or terms that co-occur more often than would be expected by chance
  - “study was to”, DC: 592,752, WC:593,718
  - “undergoing cardiac surgery”, DC: 2,589, WC: 3,135
  - “adverse cardiac”, DC: 4,405, WC:5,725

  - “in the house”, DC: 1,170, WC: 1,298
  - in house|adj|positive|E0555310, DC: 1,681, WC: 2,129

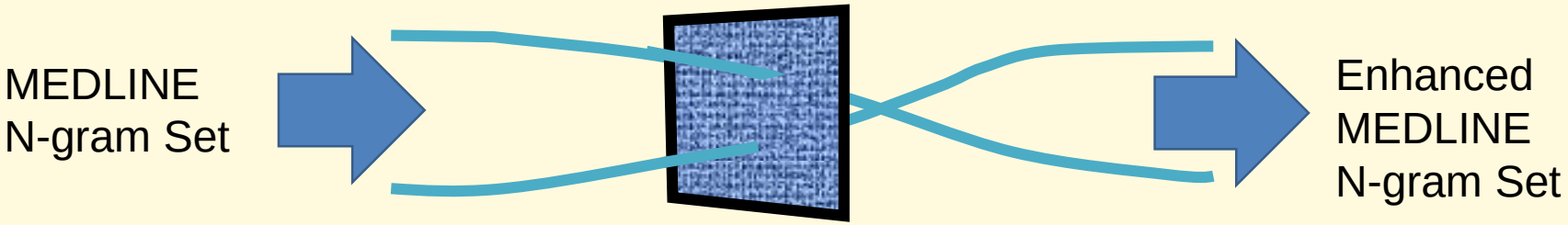
LMW Candidates from N-gram Set

- Parenthetic Acronym Pattern
  - computed tomography (CT)
  - magnetic resonance imaging (MRI)
  - polymerase chain reaction (PCR)
  - …
- EndWord Pattern
  - Syndrome: migraine syndrome, contiguous gene syndrome, …
  - Center: Heart Information Center, Veteran's Affairs Medical Center,…
  - Disease: Fabry disease, Devic disease, …
  - …
- CUI from the Metathesaurus (STMT)
  - A LMW candidate should has CUI(s)
- Spelling Variant Pattern (use distilled n-gram set)
  - SpVar normalization
  - MES (Metaphone, Edit Distance, Sorted Distance)
  - ES (Edit Distance and Sorted Distance)
- Combination of above, etc.

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

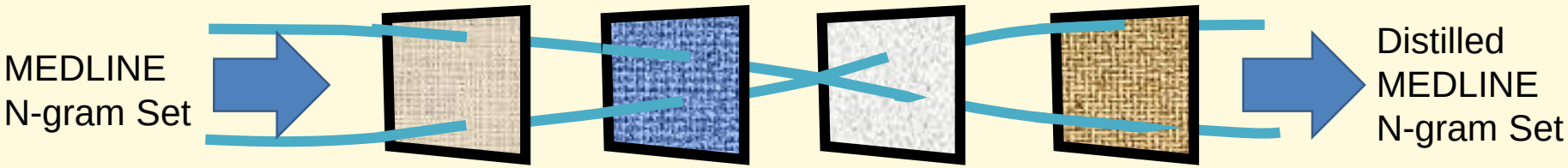
Filter



	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

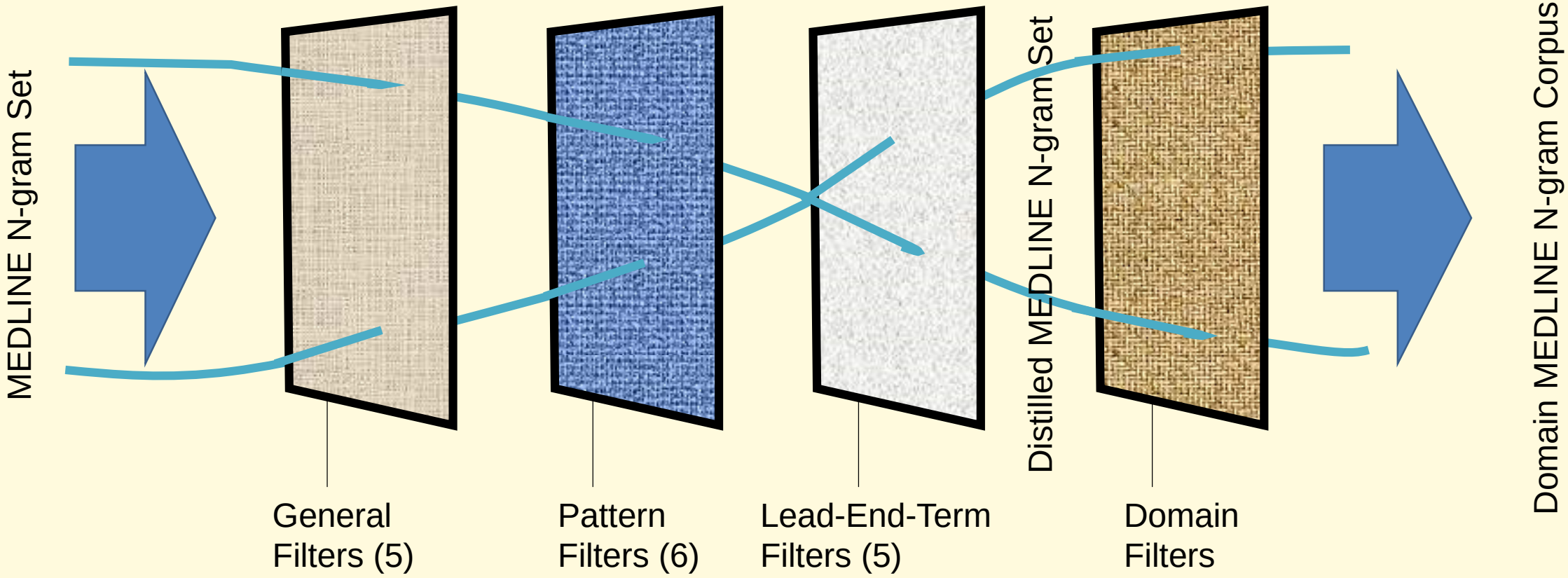
Serial Filters (High Accuracy)



	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 $_n = V_n / (V_n + I_n)$
  $= V_0 / (V_0 + I_n)$ ….. 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 (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

(~ 37.4%)

Core-term

- Strip initial and/or final punctuation from n-grams by coreterm normalization
  - Strip initial chars if they are punctuation except for closed brackets
  - Strip final chars if they are punctuation except for closed brackets
  - Recursively strip close brackets of (), [], {}, <> at both ends
  - trim

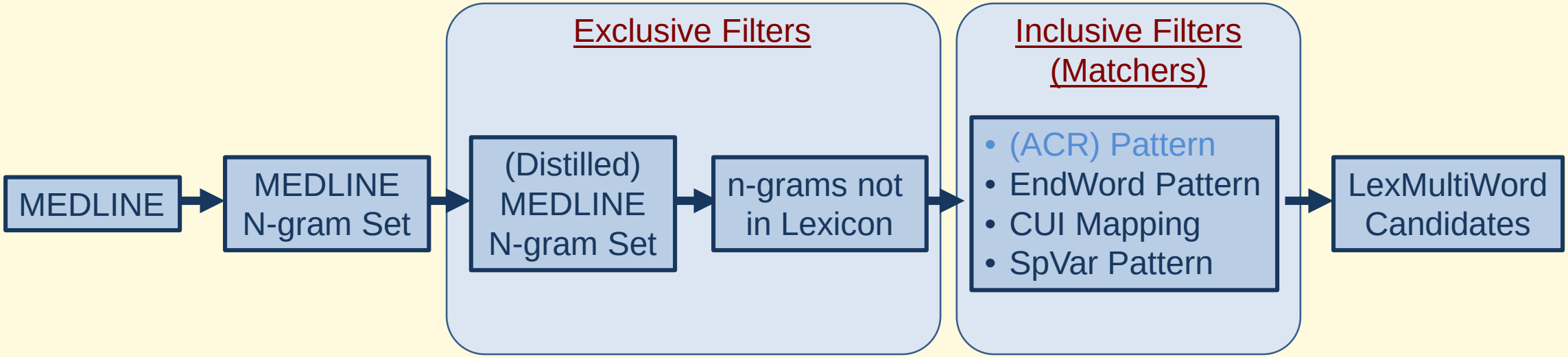
Input n-gram	Core-term
-in details	in details
in details:	in details
(in details:)	in details
(in details:))	in details:)
(-(in details)%^)	in details
{in (5) days},	in (5) days
((clean room(s)))	clean room(s)

Summary

- Distributed the MEDLINE n-gram set (2014+) to public
- Enhanced to the Distilled MEDLINE n-gram set (2014+)
- All exclusive 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
  - used as baseline for further analysis

Project Domain Exclusive Filters

Filter	Accuracy (875,890)	Pass Rate N-gram Set	Accumulated Pass Rate	Trapped Examples
Lexicon	100.0000% (0)	91.0478% (568,592)	33.9689%	• 12532576|44859301|to • 44|44|systematic name



Future Work

- Optimized Matchers:
  - Parenthetic Acronym Pattern
    - computed tomography (CT)
  - EndWord Patterns
    - Syndrome: migraine syndrome, contiguous gene syndrome, ...
  - CUI Metathesaurus
    - LMW candidate if a term has CUI(s)
  - Spelling Variant Patterns (use Distilled N-gram Set)
    - SpVar normalization
    - MES (Metaphone, Edit Distance, Sorted Distance)
    - ES (Edit Distance and Sorted Distance)
  - Combination of above to generate better LMW candidate list

Questions

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