Generating a Distilled N-Gram Set

Effective Lexical Multiword Building in the SPECIALIST Lexicon

Presented By: Dr. Chris J. Lu

2017.02.22

(HEALTHINF, Session-6)

- Lexical Systems Group: http://umlslex.nlm.nih.gov
- The SPECIALIST NLP Tools: http://specialist.nlm.nih.gov
- Chris Lu (E): chlu@mail.nih.gov

Table of Contents

1. Introduction

The SPECIALIST Lexicon

- Lexicon: A fancy synonym for “dictionary”*
- The SPECIALIST Lexicon:
  - A large syntactic lexicon of biomedical and general English
  - Designed and developed to provide lexical information needed for the SPECIALIST Natural Language Processing (NLP) system
  - Distributed in the Unified Medical Language System (UMLS) Knowledge Sources by the National Library of Medicine (NLM)

* The Insomniac’s Dictionary, Ivy Books, 1986, by Paul Hellweg, p. 140

Lexicon - Lexical Records

- POS (Part-of-Speech)
- Morphology
  - Inflection
  - Derivation
- Orthography
  - Spelling variants
- Syntax
  - Complementation for verbs, nouns, and adjectives
- Other
  - Expansions of abbreviations and acronyms
  - Nominalizations
  - …

Lexical Record Example-1

```
{base=seasickness
spelling_variant=sea sickness
spelling_variant=sea-sickness
entry=E0054880
        cat=noun
        variants=uncount
        compl=pphr(in,np)
        nominalization_of=seasick|a
dj|E0054879
}
```

Lexical Information	Example-1
Base	• seasickness
Part of speech	• noun
Inflectional morphology	• seasickness|singular
Orthography (spelling variants)	• sea sickness • sea-sickness
Abbreviation/Acronym	• N/A
Syntax (complementation)	• compl=pphr(of,np)
…	…
Derivational morphology	• **seasick|adj**
LexSynonyms	• **seasick** • **mal de mer**

Lexical Record Example-2

```
{base=seasick
spelling_variant=sea sick
spelling_variant=sea-sick
entry=E0054879
        cat=adj
        variants=inv;periph
        position=attrib(1)
        position=pred
        stative
        nominalization=seasickness|
noun|E0054880
}
```

Lexical Information	Example-2
Base	• seasick
Part of speech	• adj
Inflectional morphology	• seasick|positive
Orthography	• sea sick • sea-sick
Abbreviation/Acronym	• N/A
Syntax (complementation)	• N/A
…	…
Derivational morphology	• **seasickness|noun**
LexSynonyms	• **mal de mer|noun** • **seasickness|noun** • **…**

Lexical Record Example-3

```
{base=mal de mer
entry=E0432198
        cat=noun
        variants=uncount
}
```

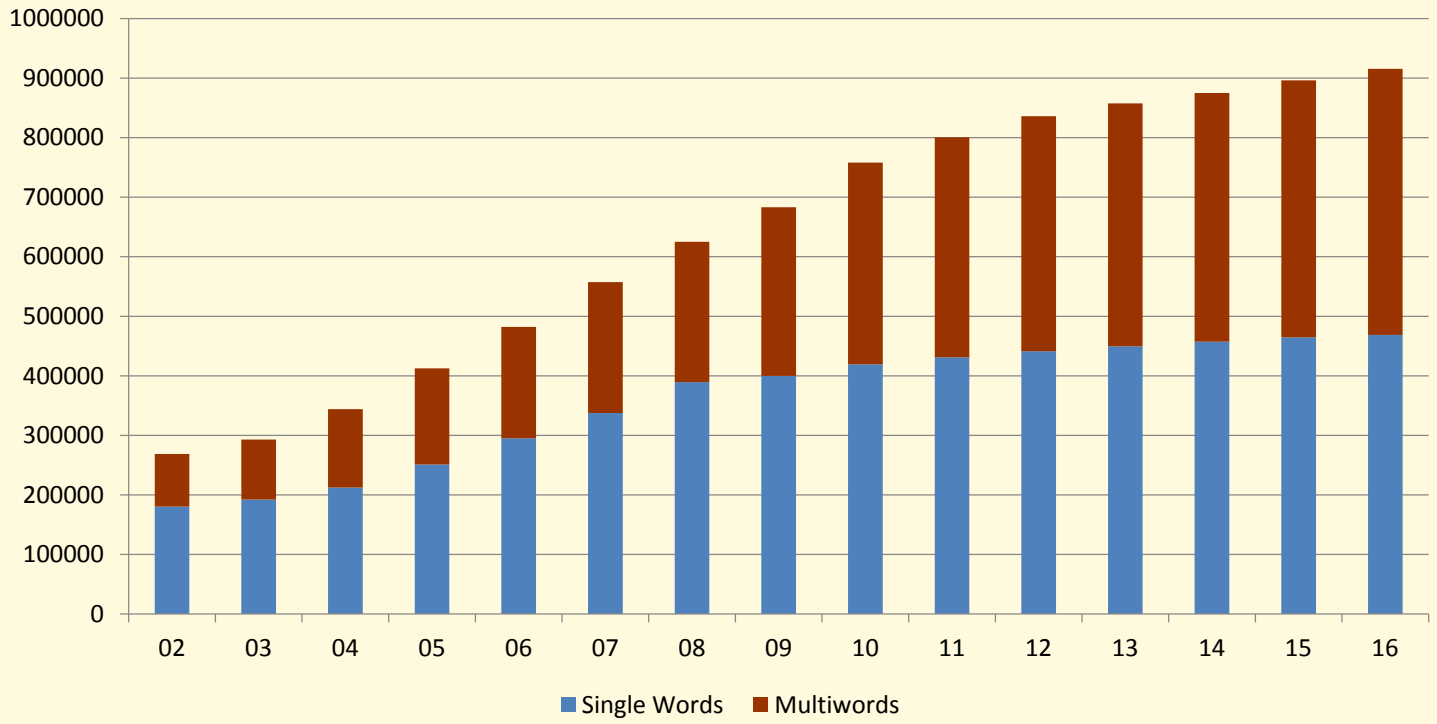
Lexical Information	Example-3
Base	• mal de mer
Part of speech	• noun
Inflectional morphology	• mal de mer|singular
Orthography	• N/A
Abbreviation/Acronym	• N/A
Syntax (complementation)	• N/A
…	…
Derivational morphology	• **N/A**
LexSynonyms	• **sea sick|adj** • **sea sickness|noun** • **…**

Lexical Records Example 1-2-3

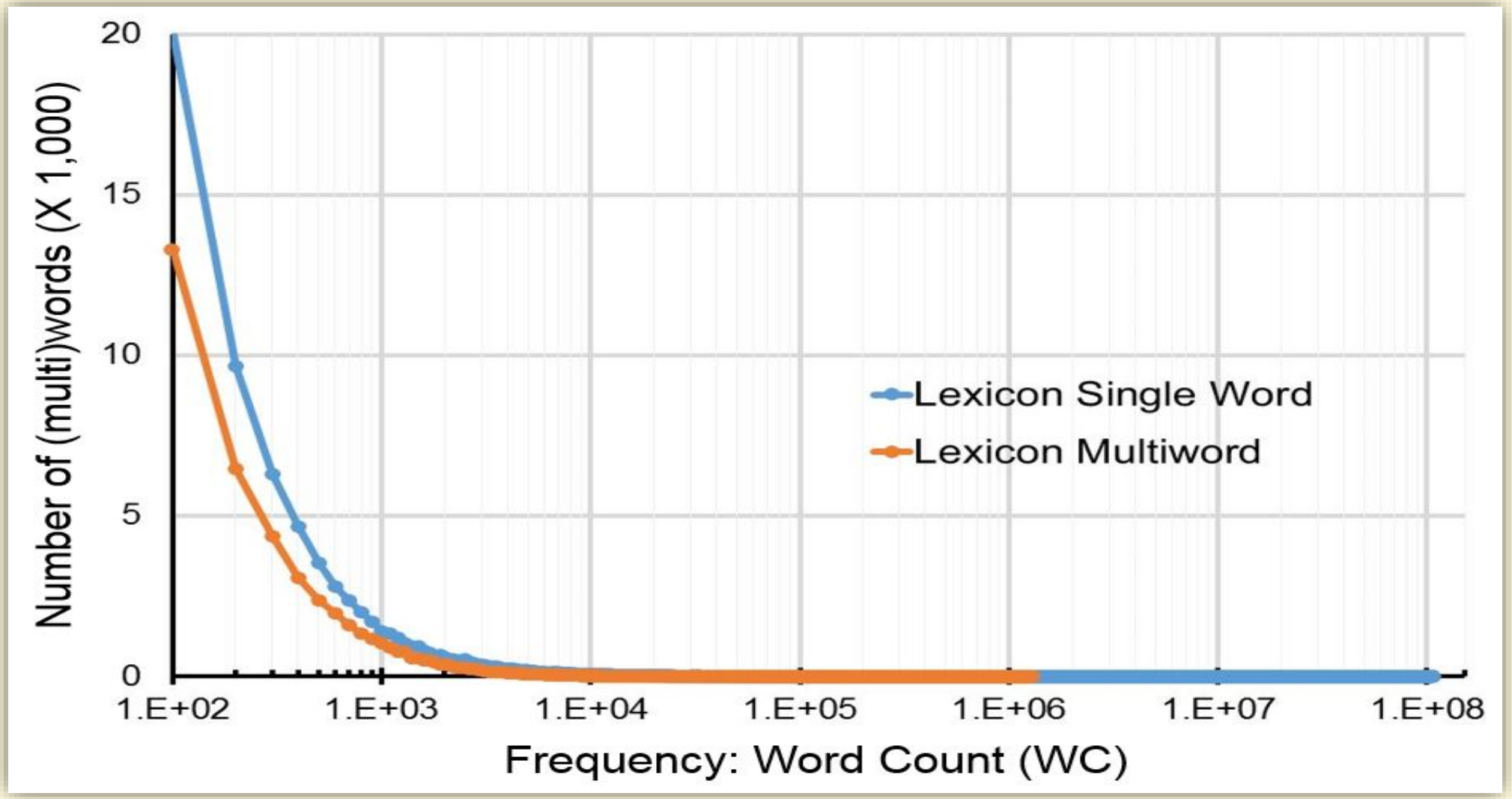
Lexical Information	Example-1	Example-2	Example-3
EUI	• E0054880	• E0054879	• E0432198
Base form of the (multi)word	• seasickness	• seasick	• mal de mer
Part of speech	• noun	• adj	• noun
Inflectional morphology	• seasickness|singular	• seasick|positive	• mal de mer|singular
Orthography	• sea sickness • sea-sickness	• sea sick • sea-sick	• N/A
Abbreviation/Acronym	• N/A	• N/A	• N/A
Syntax (complementation)	• compl=pphr(of,np)	• N/A	• N/A
…	…	…	…
Derivational morphology	• **seasick|adj**	• **seasickness|noun**	• **N/A**
LexSynonyms	• **seasick|adj** • **mal de mer|noun**	• **mal de mer|noun** • **seasickness|noun** • **…**	• **sea sick|adj** • **sea sickness|noun** • **…**

Lexicon Growth – 2002 to 2016

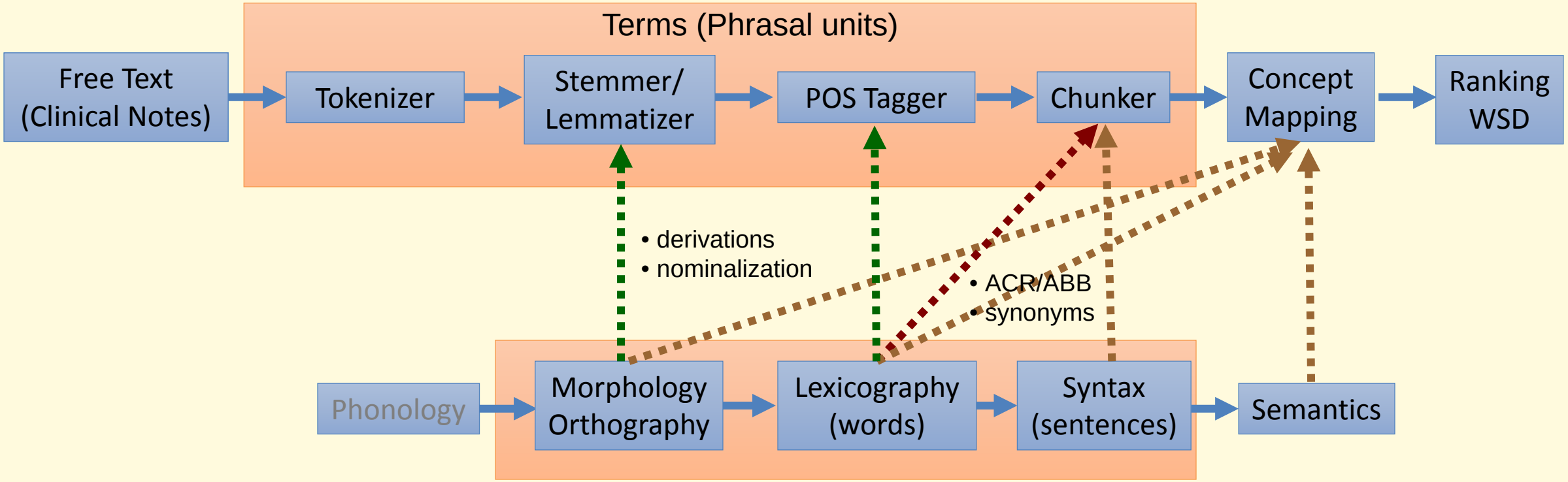
- 491,639 lexical records
- 1,090,050 words (categories and inflections)
- 915,583 forms (spelling only)
  - Single words: 468,655 (51.19%); Multiwords: 446,928 (48.81%)



The Frequency Spectrum of Lexicon on MEDLINE



NLP Pipeline



Words and Multiwords

- Lexicon terms: single words and multiwords
  - Space(s): ice-cream vs. ice cream
- Four criteria for Lexicon terms:
  - Part of Speech (POS):
    - tear break up time, frog erythrocytic virus, cardiac surgery
  - Inflection morphology (uninflection):
    - left pulmonary veins (“left pulmonary vein” and “leave pulmonary vein”)
  - Specific meaning:
    - hot dog (high temperature canine?)
  - Word order:
    - trial and error, up and down (vs. food and water, pain and fever)
    - exercise training vs. training exercise (military)

Criterion 1 - POS

- Example (PMID 9510650, TI):
  The changes of tear break up time after myopic excimer laser photorefractive keratectomy

- **Single Word Approach:**
  - Shallow Parser (U of Illinois):

    The changes of tear break up time after myopic excimer laser photorefractive keratectomy

    NP PP NP VP PRT NP PP NP

  - Parser (Stanford):

    The changes of tear break up time after myopic excimer laser photorefractive keratectomy

    NP PP NP VP PRT NP SBAR NP

- **Multiword Approach:**
  - The changes of tear break up time after myopic excimer laser photorefractive keratectomy

    E0635418|Noun → C2111106 Lacrimation tear break-up time

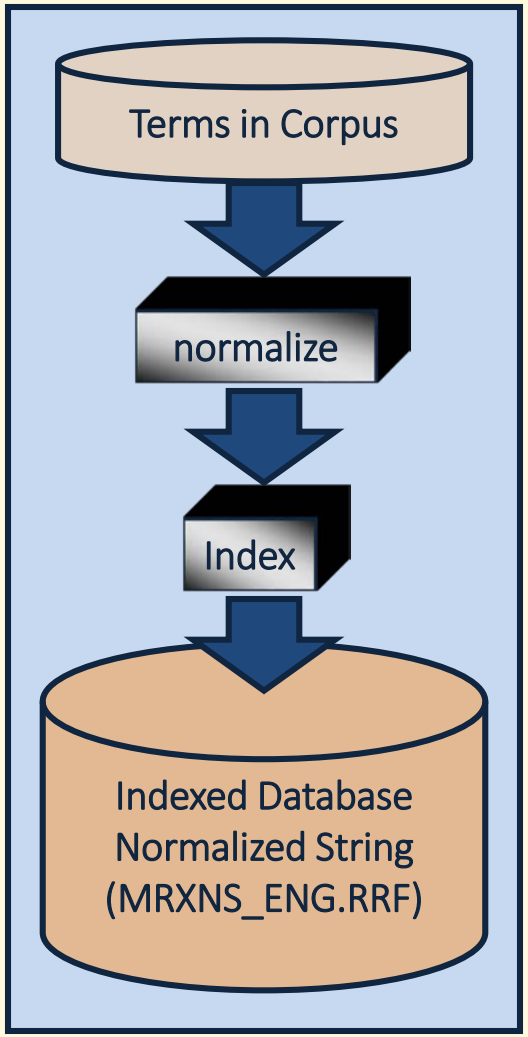
    E0764487|Noun → C2069669 excimer laser photorefractive keratectomy

Criterion 2 – Norm (Uninflection)

- 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
- …

disease hodgkin

- C0019829
- Hodgkin Disease



Criterion 2 – Norm (Uninflection)

Hodgkin's Diseases, Diseases of Hodgkin, …

Query

norm

Normed Term

disease hodgkin

SQL

Indexed Database
Normalized String

Results that matches the normalized query
- C0019829
- Hodgkin Disease

Criterion 2 – Inflection/Uninflection

➢ Example (PMID 14479709, TI):
Complete anomalous drainage of left pulmonary veins.

- **Single Word approach** (norm - uninflection):
  - left pulmonary vein
  - **leave pulmonary vein**

- **Multiword Approach**:
  - left pulmonary vein

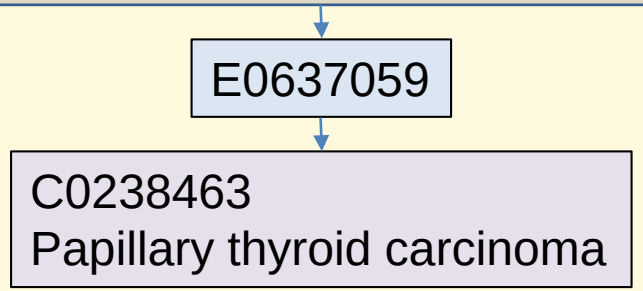
Inflectional Variants	Adjective (Positive)	Adverb (Positive)	Noun (Singular)	Verb (Infinitive)
left	left	left	left	leave
pulmonary	pulmonary			
veins			vein	vein
left pulmonary veins			left pulmonary vein	

Criterion 3 – Meaning: Name Entity Recognition (NER)

- Example (PMID 23477346, TI) – Lead-term:
  - Follicular variant of papillary thyroid carcinoma is a unique clinical entity.

- Example (PMID 581461, TI) – End-term:
  - Nucleolar abnormalities in human papillary thyroid carcinoma.

- Example (PMID 6143549 , AB) – Mid-term:
  - Coexisting papillary thyroid carcinoma occurred in three patients with HCA.



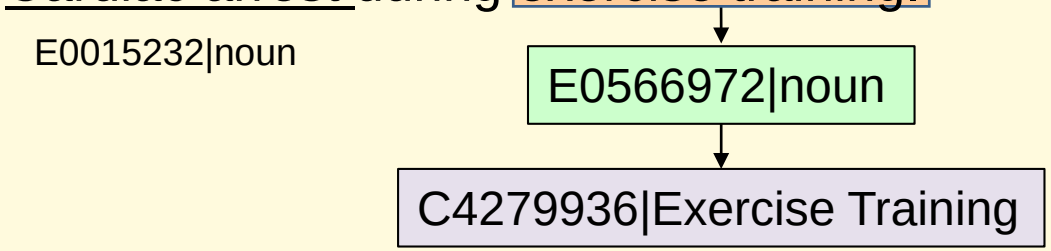
Criterion 3 – Meaning: Overlap Window

- Example (PMID 12792778, TI):
- Inhibition of metastatic brain tumor growth by intramuscular administration of the endostatin gene.
  - Lead-Term: metastatic brain tumor growth: C0220650
  - End-Term: metastatic brain tumor growth: C0598934
  - LMWs: metastatic brain tumor|C0220650, brain tumor|C0006118, tumor growth|C0598934, ..

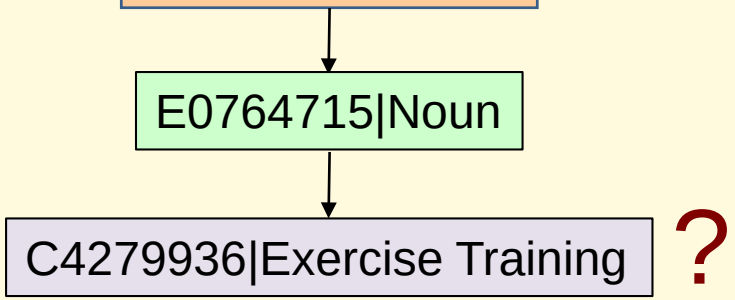
- Example (PMID 20162874, AB):
- In the present patient right pulmonary agenesis is co-occurring with VACTERL syndrome.
  - Lead-Term: the present patient right pulmonary agenesis: C0030706
  - End-Term: the present patient right pulmonary agenesis: C0265784
  - LMWs: patient right|C0030706, right pulmonary agenesis|C0265784, pulmonary agenesis|…

Criterion 4 – Word Order (Norm)

- Example (PMID 5820369, TI):
  Cardiac arrest during exercise training.



- Example (PMID 14719633, AB):
  Military training exercises are conducted routinely in the Mojave Desert.



LMWs and MWEs

- LMWs are a subset of MWEs (Multiword Expressions)
  - MWEs that have specific inflection morphology, POS, meaning, and word order are LMWs (except for complementation and idioms)

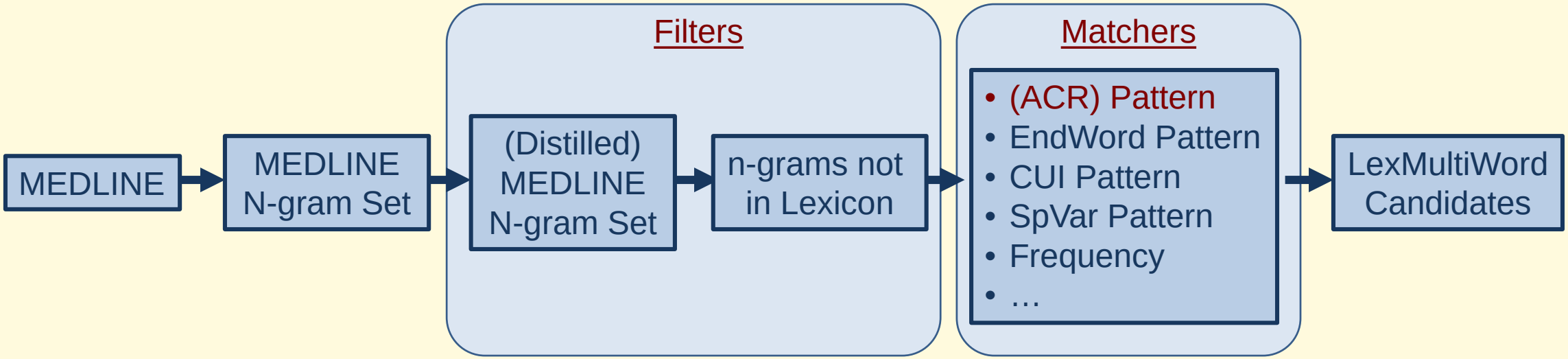
	Descriptions	Examples
LMWs and MWE (Overlap)	• Phrasal position • Adverb, adjective • Fixed phrases (non-decomposable)	• because of, due to • face down, in house • kingdom come, by and large
LMWs and MWE (Difference)	• Collocation • Specific meaning • Complementation • Idioms • …	• undergoing cardiac surgery • in the house • beat someone up, give birth • kick the bucket, shoot the breeze

Project Objective

- Effective lexical multiword building
  - Develop a systematic approach
  - Add multiwords from the latest MEDLINE to the SPECIALIST Lexicon
  - Generate high precision multiword candidate list

2. Approach

LMWs - Processes



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

* 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

- Five-grams: covers 99.38% multiwords in Lexicon.2016
- Max. Word Length: 50 to cover 99.46% words in Lexicon.2016

The MEDLINE N-gram Set

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

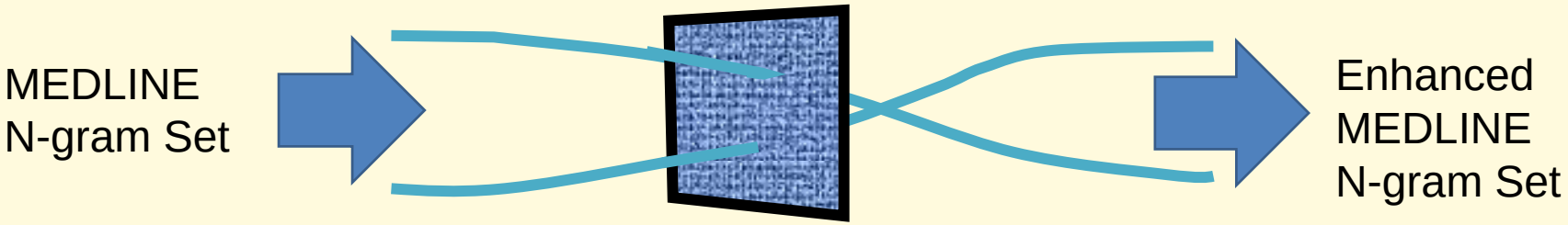
N-grams	2014	2015	2016
unigrams	804,382	843,206	883,287
bigrams	4,587,349	4,845,965	5,114,547
trigrams	6,287,536	6,702,194	7,134,807
four-grams	3,799,377	4,082,612	4,380,474
five-grams	1,545,175	1,674,715	1,812,223
n-gram set	17,023,819	18,148,692	19,325,338

Enhancing N-gram Set

- 17 ~ 19 million is a big number
  - Includes many n-grams that are invalid lexical terms
    - Ex: “increased significantly|78545”
  - Too big for further processes by computer programs with complicated algorithm

- Reduce the size by filtering out invalid multiwords:
  - increase precision
  - without sacrificing recall
  - the distilled MEDLINE n-gram set

Exclusive Filter

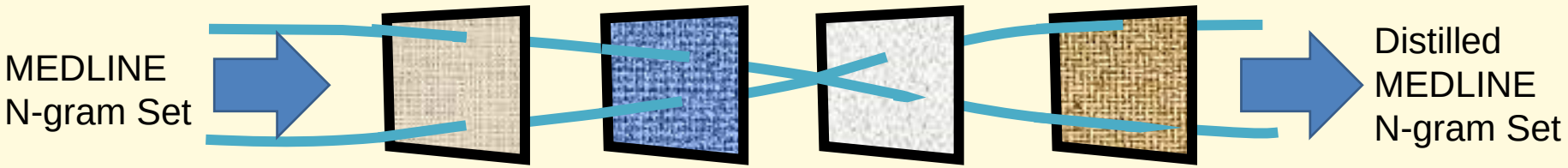


	Trap (not retrieved)	Pass (retrieved)
Valid (relevant)	FN	TP
Invalid (not relevant)	TN	FP

- Filter passing rate = pass-through terms / total terms
- Recall = TP / TP + FN

- Recall Test: apply filters on Lexicon.2016 (valid 915,583 word set where TN & FP are 0)
  - Recall = TP / TP + FN
    = pass-through terms / total terms
    = passing rate

Serial Filters (High Recall)



	N-gram	Filter-1	Filter-2	…	Filter-N	Distilled
Valid (TP)	TP_0	TP_1	TP_2	…	TP_x	TP_x
Invalid (FP)	FP_0	FP_1	FP_2	…	FP_x	FP_x

- Applied filters with high recall rate ($TP_0 \cong TP_1 \cong \ldots \cong TP_x$; $FP_0 > FP_1 > \ldots > FP_n$)
- N-gram Precision $_x = TP_x / (TP_x + FP_x)$
  $\cong TP_0 / (TP_0 + FP_x)$ ….. TP_x is similar to TP_0 (high recall)
  $> TP_0 / (TP_0 + FP_0)$ ….. FP_0 is bigger than FP_n (high recall)
  $>$ Precision $_0$
- N-gram Recall $_x = TP_x / (TP_x + FN_x)$
  $= TP_x / (TP_x + FN_0)$ ….. FN_n is same as FN_0, a constant
  $\cong TP_0 / (TP_0 + FN_0)$ ….. TP_n is similar to TP_0 (high recall)
  $\cong$ Recall $_0$

Independent High Recall Filters

Filter Type	High Recall Filter
General Exclusive Filters (5)	• Pipe • Punctuation or space • Digit • Number • Digit and Stopword
Pattern Exclusive Filters (6)	• Parenthetic acronym - (PAP) • Indefinite Article • UPPERCASE Colon • Disallowed Punctuation • Measurement • Incomplete
Lead-End-Term Exclusive Filters (5)	• Absolute Invalid Lead-Term • Absolute Invalid End-Term • Lead-End-Term • Lead-Term No SpVar • End-Term No SpVar

1. General Exclusive Filters

Filter	Recall Test* On Lexicon (Trapped)	Passing Rate On MNS (Trapped)	Cumulative Passing Rate On MNS	Trapped Examples
Pipe	100.0000% (0)	100.0000% (7)	100.0000%	• 38|44|(|r| • 33|37|Ag|AgCl
Punctuation or space	100.0000% (0)	99.9977% (425)	99.9978%	• 1259147|3690494|= • 604567|2377864|+/-
Digit	99.9999% (1)	99.3136% (132,650)	99.3114%	• 1404799|2062240|2 • 239725|499064|95%
Number	99.9953% (41)	99.9775% (4,326)	99.2890%	• 2463066|3359594|two • 18246|20674|first and second
Digit and Stopword	99.9991% (8)	99.1777% (157,786)	98.4725%	• 3155416|4125616|on the • 11180|12722|1, 2, and

* Recall test on Lexicon.2016: 915,583 words

2. Pattern Exclusive Filters

Filter	Recall Test* On Lexicon (Trapped)	Passing Rate On MNS (Trapped)	Cumulative Passing Rate On MNS	Trapped Examples
Parenthetic acronym - (PAP)	100.0000% (0)	98.9647% (197,022)	97.4530%	• 33117|33381|chain reaction (PCR) • 30095|30315|polymerase chain reaction (PCR)
Indefinite Article	99.9986% (13)	98.1713% (344,403)	95.6709%	• 270384|292590|a case • 40271|40512|A series
UPPERCASE Colon	99.9999% (1)	99.3838% (113,936)	95.0813%	• 2069343|2070116|RESULTS: • 18015|18016|AIM: The
Disallowed Punctuation	99.9986% (13)	99.2625% (135,508)	94.3801%	• 324405|719011|(n = • 86525|133350|(P < 0.05)
Measurement	99.9920% (73)	98.1572% (336,112)	92.6409%	• 154905|181001|two groups • 12160|15197|10 mg/kg
Incomplete	100.0000% (0)	99.0708% (166,356)	91.7801%	• 482021|1107869|(P • 25347|25992|years) with

* Recall test on Lexicon.2016: 915,583 words

3. Lead-End-Terms* Exclusive Filters

Filter	Recall Test* On Lexicon (Trapped)	Passing Rate On MNS (Trapped)	Cumulative Passing Rate On MNS	Trapped Examples
Absolute Invalid Lead-Term	99.9943% (52)	73.4329% (4,712,162)	67.3967%	• 2780043|3451203|of a • 432921|434591|this study was
Absolute Invalid End-Term	99.9997% (3)	79.1897% (2,710,470)	53.3713%	• 1878109|3534031|patients with • 1062545|1261445|between the
Lead-End-Term	99.9992% (7)	99.9739% (2,687)	53.3573%	• 2578756|3106139|in a • 1733|1744|For one
Lead-Term No SpVar	**99.9913% (80)**	85.9342% (1,450,394)	45.8522%	• 658430|708246|to determine • 533913|554628|In addition,
End-Term No SpVar	99.9968% (29)	83.5433% (1,458,246)	**38.3064%**	• 1009451|1295670|number of • 726|734|(HPV) in

Cum. Recall: 99.96%

* Recall test on Lexicon.2016: 915,583 words

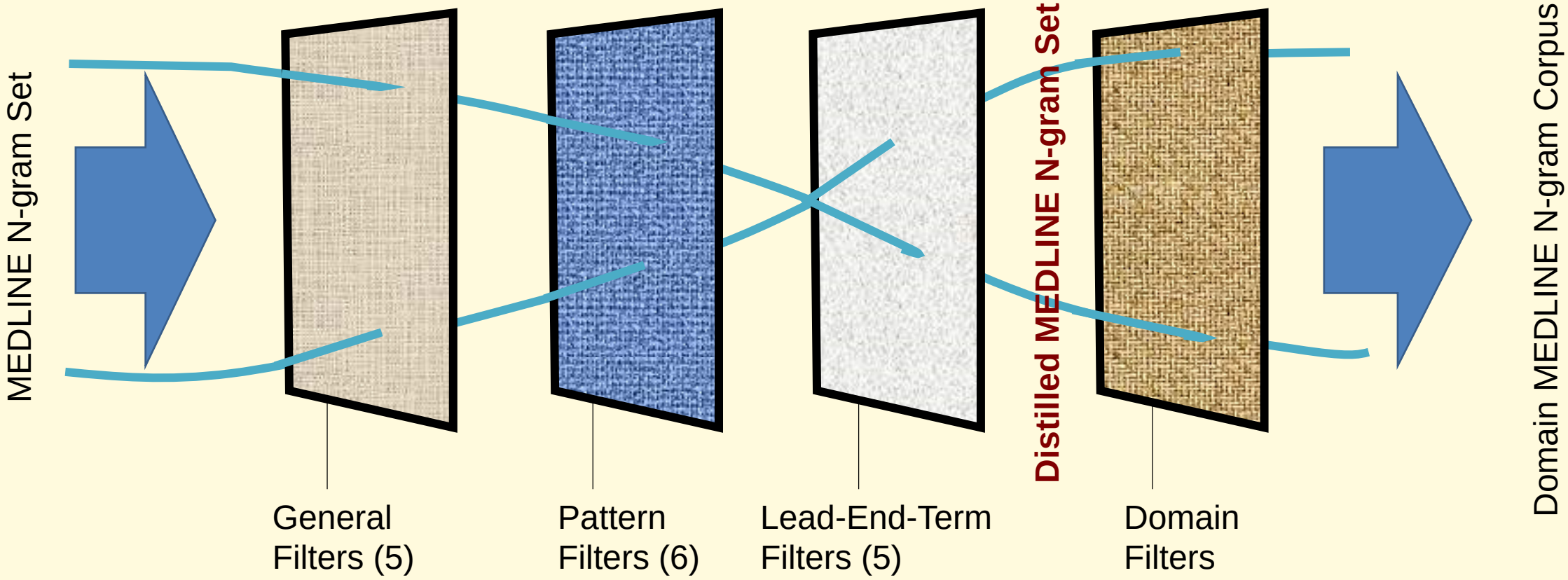
** Lead-end-terms: derived from auxiliaries (be, do), complementizer (that), conjunctions (and, or, but) determiners (a, the, some), modals (may, must, can), pronoun (it, he, they), preposition (to, on, by)

The Distilled MEDLINE N-gram Set (DMNS)

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

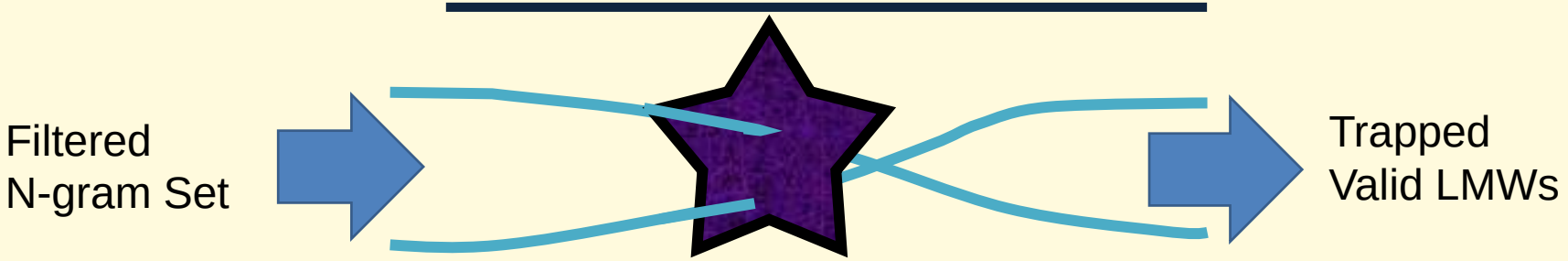
N-grams	2014	2015	2016
unigrams	804,382	843,206	883,287
bigrams	4,587,349	4,845,965	5,114,547
trigrams	6,287,536	6,702,194	7,134,807
four-grams	3,799,377	4,082,612	4,380,474
five-grams	1,545,175	1,674,715	1,812,223
N-gram Set	17,023,819	18,148,692	19,325,338
Distilled N-gram Set	6,351,392	6,793,561	7,402,848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

Distilled N-gram Set



3. Tests and Results

Test: PAP Matcher



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

- Parenthetic Acronym Pattern (PAP) Matcher

Process	Example-1	Example-2	Example-3
1. Apply PAP Filter	clear cell sarcoma (CCS)	cell sarcoma (CCS)	Unified Health System (SUS)
2. Retrieve expansion	clear cell sarcoma	cell sarcoma	~~Unified Health System~~
3. Remove invalid expansion	clear cell sarcoma	~~cell sarcoma~~	

Evaluation of Distilled MEDLINE N-gram Set

- Test on MNS and DMNS
  - generate 17,707 LMW candidates from PAP matchers on MNS.2016,
  - manually tagged and reviewed by 3 linguists: TP (15,850), FP (1,857)

Case	Test Case - Model	TP	FP	FN	TN	Precision	Recall	F1
1	**PAP matcher on MNS (baseline)**	15,850	1,857	0	0	0.8951	(1.0000)	0.9447
2	**PAP matcher on DMNS (16 filters)**	15,840	1,299	10	558	0.9242	0.9994	0.9603

- Similar results on the recall test and PAP matcher test for 2014-2016

	2014	2015	2016
Recall test: cumulative passing rate	37.31%	37.43%	38.30%
Recall test: cumulative recall (product)	99.97%	99.97%	99.96%
PAP Matcher Test: recall	99.96%	99.94%	99.94%
PAP Matcher Test: precision	94.31%	91.89%	92.42%

4. Applications

Spelling Variant Pattern (SVP) Matcher

- Utilize SpVar Norm, Metaphone, edit distance, sorted distance, etc.

Step	Algorithm	ED	Examples	SpVarNorm	Metaphone	ED	SD	Recall On Lexicon	Run Time
1	SpVar Norm	N/A	• *St. Anthony's fire* • *Saint Anthony's fire* • *lamin-A* • *lamin A*	• *Saint Anthony's fire* • *Lamin A*				0.8050	1 min.
2	MES	2	• yuppie flu • yuppy flu	• *yuppieflu* • *yuppyflu*	• [YPFL]	2	✓	0.9792	7 hr
3	ES	1	• *zincemia* • *zincaemia*	• *zincemia* • *zincaemia*	• [SNSM] • [SNKM]	1	✓	0.9931	23 hr
4	MES	3	…	…	…	…	…	0.9940	8 min
5	ES	2	…	…	…	…	…	0.9970	26 hr
6	MES	4	…	…	…	…	…	0.9972	2 min

56 hr

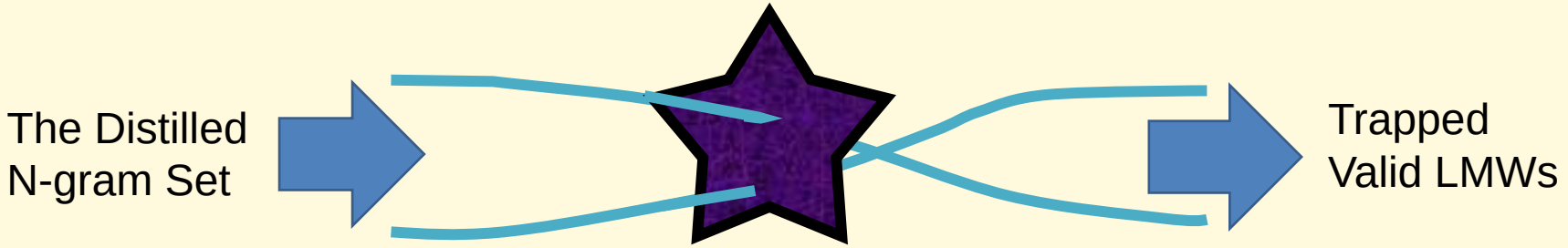
* Lexicon: 0.9 M (56 Hr); MNS 20 M (too long); DMNS 7.4 M (20 days)

Practice Results

- Baseline: 17,707 LMW Candidates from (ACR) matcher, tagged by linguists

Case	Test Case - Model	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
1	**PAP matcher on MNS (baseline)**	15,850	1,857	0	0	0.8951	(1.0000)	0.9447	0.8951
2	**PAP matcher on DMNS (16 filters)**	15,840	1,299	10	558	0.9242	0.9994	0.9603	0.9261
3	**SpVar pattern matcher on case 2**	8,094	499	7,756	1,358	0.9419	0.5107	0.6623	0.5338

Other Matchers



- 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.
  - 33 high frequency end-terms
- Frequency Strategy
- Combination of above models

Practice Results

- Baseline: 17,707 LMW Candidates from (ACR) matcher, tagged by linguists

Case	Test Case - Model	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
1	**PAP matcher on MNS (baseline)**	15,850	1,857	0	0	0.8951	(1.0000)	0.9447	0.8951
2	**PAP matcher on DMNS (16 filters)**	15,840	1,299	10	558	0.9242	0.9994	0.9603	0.9261
3	**SpVar pattern matcher + Distilled**	8,094	499	7,756	1,358	0.9419	0.5107	0.6623	0.5338
4	**Metathesaurus CUI pattern matcher**	10,056	755	5,794	1,102	0.9302	0.6344	0.7544	0.6301
5	**EndWord pattern matcher (Top 33)**	2,346	251	13,504	1,606	0.9034	0.1408	0.2544	0.2232
6	**Distilled + CUI + SpVar**	5,892	212	9,958	1,645	0.9653	0.3717	0.5368	0.4257
7	**Distilled + CUI + SpVar + EndWord (33)**	992	15	14,858	1,842	0.9851	0.0626	0.1177	0.1600
8	**Distilled + CUI + EndWord (33)**	1766	113	14,084	1,744	0.9399	0.1114	0.1992	0.1982

Summary

- Developed high recall filters (> 99.99% on Lexicon recall test)
- Obtained the distilled MEDLINE n-gram set at passing rate of 37-38%
  - smaller data set
  - better precision
  - similar recall
- Applied filters and matchers to generate LMW candidate list
- Improve lexBuilding on multiwords (> 40% efficiency improvement)

- Distribute the MEDLINE n-gram set (2014+) to public
- Distribute the Distilled MEDLINE n-gram set (2014+) to public
- Generic filters for other corpora

Conclusion

- Multiwords are pervasive, challenging and vital to NLP
- Build a biomedical Lexicon to support NLP and MLP
  - POS
  - Morphology (inflections and derivations)
  - Specific meaning
  - Specific order
- Multiword Approach (embedded information)

Future Work

- Continuously develop filters and matchers for LexBuilding on multiwords
- Develop different matchers for better LMW candidate list
- Develop new SPECIALIST NLP Tools based on multiword approach

Acknowledgements

- Supported by the Intramural Research Program of the NIH/NLM

- Co-authors (NLM):
  - Destinee Tormey
  - Dr. Lynn McCreedy
  - Allen C. Browne

- Mr. Guy Divita (Utah University)

Questions

- Lexical Systems Group: http://umlslex.nlm.nih.gov
- SPECIALIST NLP Tools: http://specialist.nlm.nih.gov
- Chris Lu (E): chlu@mail.nih.gov