Multiwords

By: Dr. Chris Lu

The Lexical Systems Group

NLM. LHNCBC. CGSB

July, 2014

Table of Contents

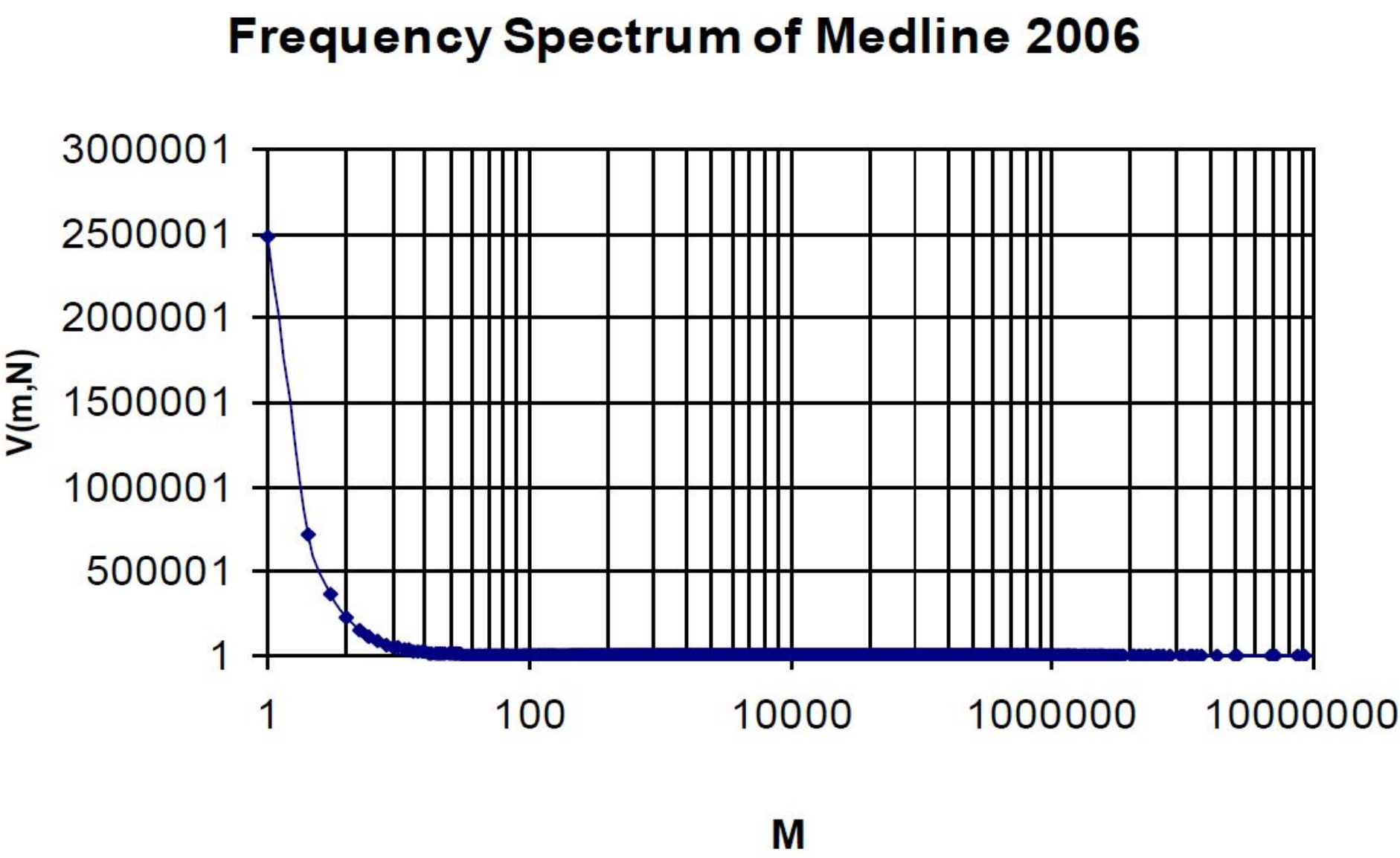
The SPECIALIST NLP Tools

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

Objective

- A systematic way to add more higher frequency words form MEDLINE to the SPECIALIST Lexicon

- A better understanding on:
  - words in MEDLINE
  - current status of Lexicon


Frequency Spectrum of Medline 2006
3000001
2500001
2000001
1500001
1000001
500001
1
V(m,N)
1
100
10000
1000000
10000000
M

What is a Word?

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

  - saw|verb|infinitive|E0054444

  - saw|verb|past|E0055007

- Word boundary – space or tab
- Single words vs. multiwords (MWEs)

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

LexBuild Process

- Built by linguists
- LexBuild: a web-based computer-aided tool
- Resources: a list of words (element words)
  - Add new lexical records if no exact/close match
  - Update existing lexical records if related records are found by close match
  - Multiwords that contain these words are reviewed through the Essie search engine*, Google Scholar, dictionaries, biomedical publications, domain-specific databases, nomenclature guidelines, and books, etc.

* N.C. Ide, R.F. Loane, D.D. Fushman, “Essie: A Concept-based Search Engine for Structured Biomedical Text”, JAMIA, Vol. 14, Num. 3, May/June, 2007, p.253-263

Element Word Approach

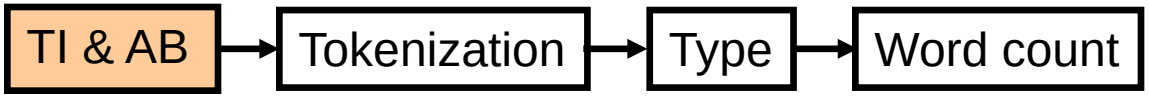
- Element words are lowercase single words without punctuation and are not stopwords

Single words/Multiwords	Element words
• saw	• saw
• ice-cream • ice cream	• ice • cream
• clubfoot	• clubfoot
• club-foot • club foot	• club • foot
• drop-foot gait	• drop • foot • gait
• Food and Drug Administration	• food • drug • administration

Stop Words

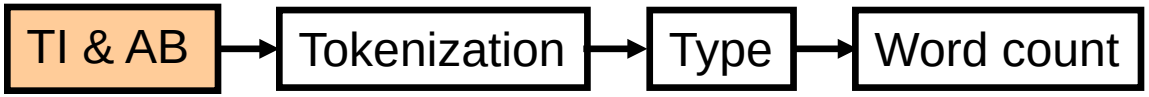
- A high frequency words - preposition
- A grammar word –not too much meaning
- Examples:
  - Lexical Tools (11): of, and, with, for, nos, to, in, by, on, the, (non mesh)
  - Text Categorization (11,068): com, edu, htm, html, www, pdf, abandon, abandoned, etc.

Element Words



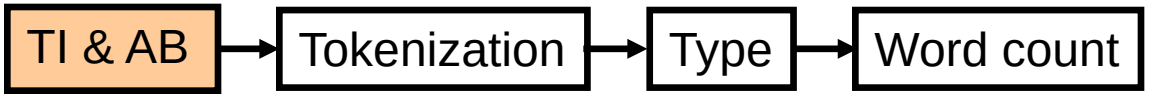
PMID- 961031
OWN - NLM
STAT- MEDLINE
DA - 19761020
DCOM- 19761020
LR - 20041117
PUBM- Print
IS - 0042-2835 (Print)
VI - 10
IP - 1
DP - 1976 Jan-Feb
TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
PG - 30-7
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, serious or not serious (100%). Ventricular premature beats were the most frequent type of arrhythmia in the first and second postoperative days (80%). Two cases expired postoperatively. In one of them complete atrioventricular block developed after double valvular replacements (mitral and tricuspid). The other died of low cardiac output syndrome. Etiology of the arrhythmias
...

Element Words



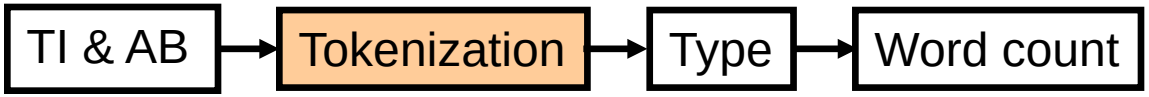
PMID- 961031
OWN - NLM
STAT- MEDLINE
…
DP - 1976 Jan-Feb
TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
PG - 30-7
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, …
…
JT - Vascular surgery
JID - 0103277

Element Words



PMID- 961031
OWN - NLM
STAT- MEDLINE
…
DP - 1976 Jan-Feb
TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
PG - 30-7
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, …
…
JT - Vascular surgery
JID - 0103277

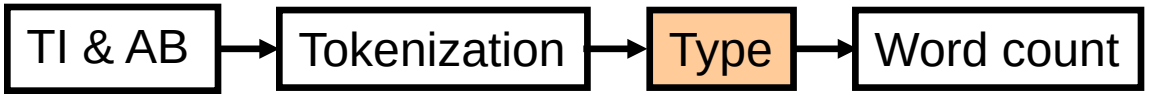
Element Words



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, …

- Utilize Lexical Tools to retrieve element words:
  - Lowercase (lvg -f:l)
  - Replace punctuation with space (lvg -f:o)
  - Tokenize (wordInd)

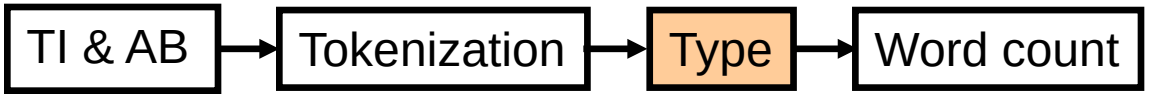
Element Words



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, …

Types	Descriptions	Example
LEXICON	Single words in Lexicon	open, surgery, etc.
NUMBER	Numbers in Lexicon	three, fifty, etc.
DIGIT	Pure digit	3, 50, 015, etc.
MULTIWORD	Not a single word, but part of multiword in Lexicon	non, vitro, mellitus, etc.
NEW	New element words, not in above types	cdh, mfi, etc.

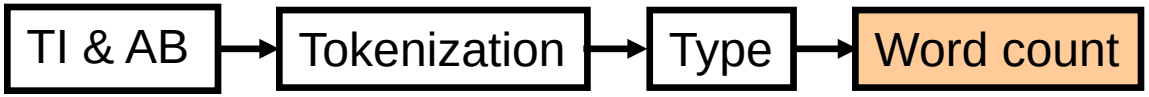
Element Words



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, …

Word	Type
postoperative	LEXICON
arrhythmias	LEXICON
in	LEXICON
open	LEXICON
heart	LEXICON
fifty	NUMBER
50	DIGIT
…	…

Element Words



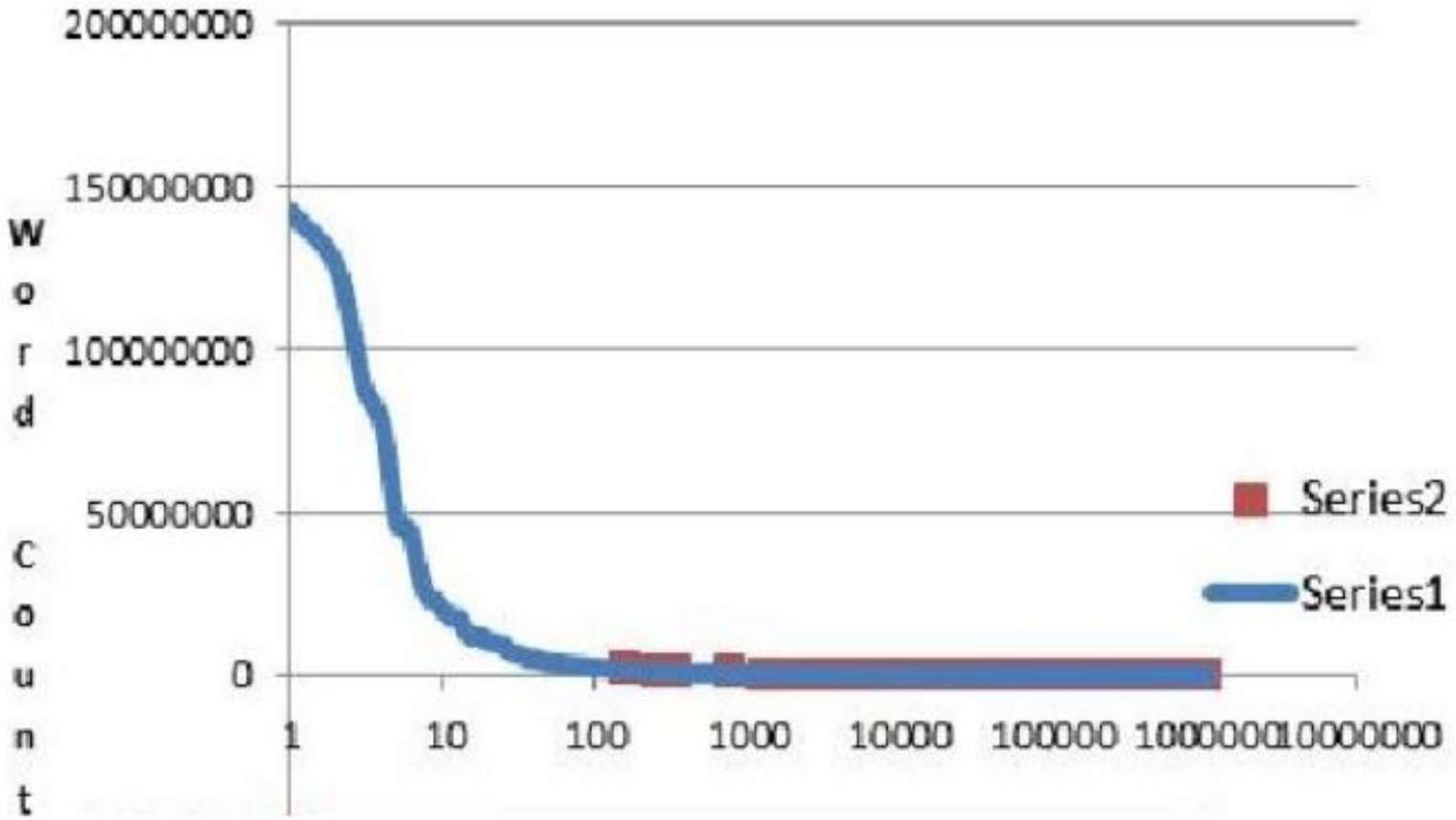
TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, …

Word	Type	WC
postoperative	LEXICON	3
arrhythmias	LEXICON	2
in	LEXICON	3
open	LEXICON	3
heart	LEXICON	2
fifty	NUMBER	1
50	DIGIT	1
…	…	…

MEDLINE - 2014

Rank	Word	Type	WC	Accu WC	Accu. %
1	the	LEXICON	142,300,698	142,300,698	5.2207%
2	of	LEXICON	125,686,598	267,987,296	9.8318%
3	and	LEXICON	89,652,562	357,639,858	13.1210%
4	in	LEXICON	77,423,109	435,062,967	15.9615%
5	to	LEXICON	46,838,419	481,901,386	17.6798%
…	…	…	…	…	…
16	1	DIGIT	11,924,150	716,284,063	26. 2788%
…	…	…	…	…	…
104	three	NUMBER	2,401,132	1,112,879,272	40.8290%
…	…	…	…	…	…
155	non	MULTIWORD	1,766,578	1,216,410,810	44.6273%
…	…	…	…	…	…
3,984	c57bl	NEW	72,178	2,282,311,101	83.7327%
…	…	…	…	…	…
3,264,205	zzzzzzzzzzzzzz	NEW	1	2,725,710,505	100.0000%

MEDLINE 2014 – F.S.



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		

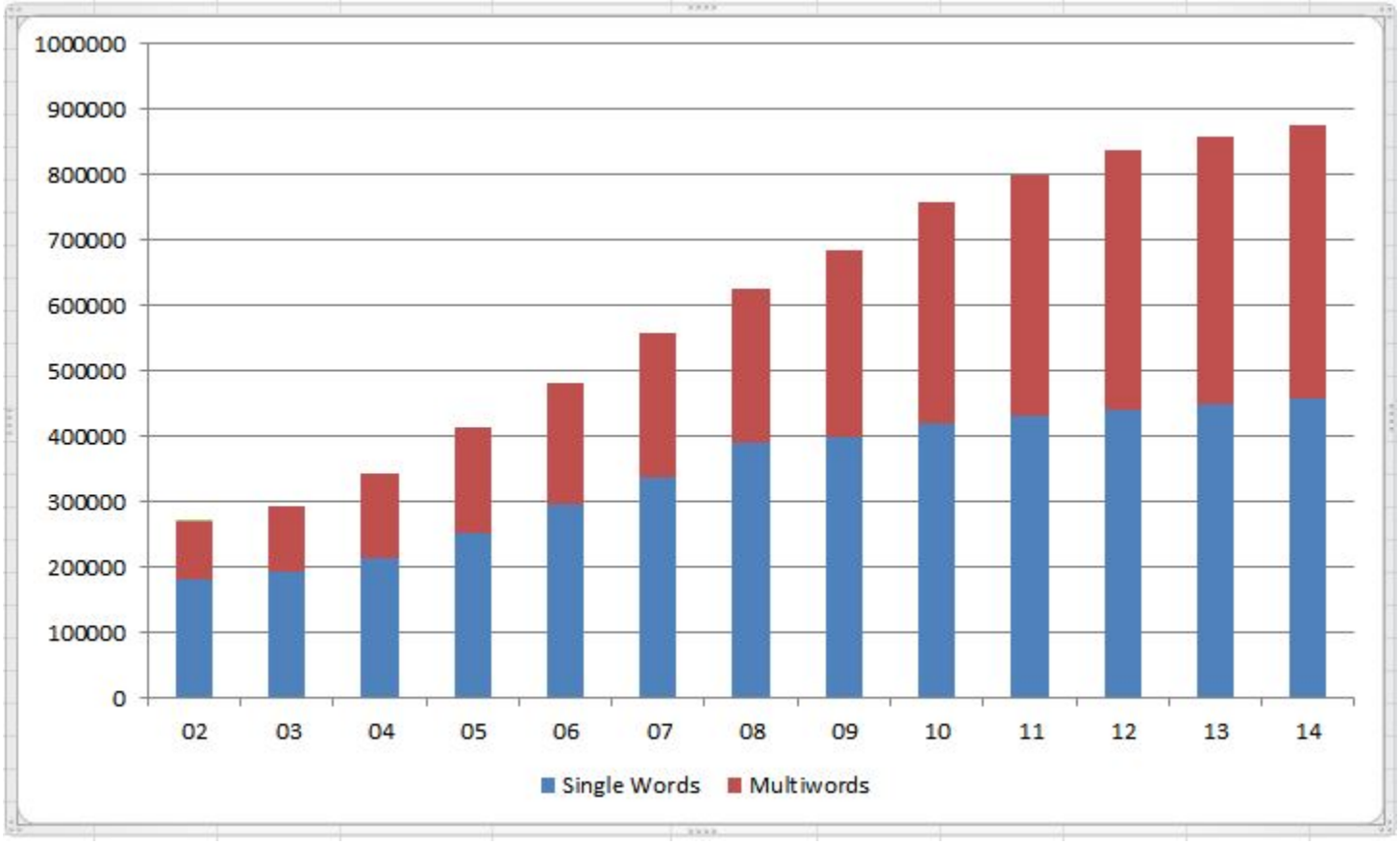
Lexicon Coverage
(by unique word - spelling)

- 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		

Lexicon.2014

- 476,857 lexical records
- 1,047,427 words (categories and inflections)
- 875,090 forms (spelling only)
  - Single words:457,335 (52.26%)
  - Multiwords: 417,755 (47.74%)



MEDLINE – 2014 (NEW)

No.	Rank	Word	Type	WC	Accu. %
1	3,984	c57bl	NEW	72178	83.7327%
2	4,161	yl	NEW	68165	84.1885%
3	4,749	h2o2	NEW	57215	85.5367%
4	5,266	cm2	NEW	49079	86.5408%
5	5,663	bax	NEW	44175	87.2185%
...	...	...	...	...	...
1,565	40,543	9b	NEW	1500	97.5841%
...	...	...	...	...	...
2,549	49,779	zw	NEW	1000	97.9990%
...	...	...	...	...	...
37,826	162,875	zwanzig	NEW	100	99.2913%
...	...	...	...	...	...
360,382	614,380	zyklophin	NEW	10	99.7746%
...	...	...	...	...	...
1,508,207	1858746	000000125x	NEW	1	99.9484%
...	...	...	...	...	...
2,855,422	3,264,205	zzzzzzzzzzzzzzz	NEW	1	100.0000%

Example – New Element Words

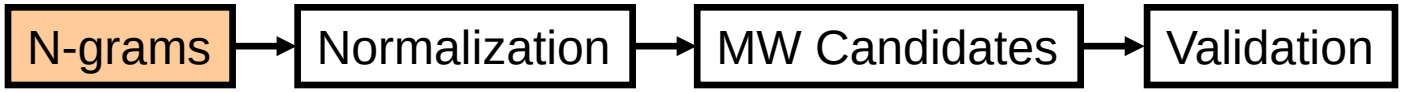
- New element words:
  - WC > 1500
  - Example: cdh|9982|93.8584%
    - 44 new lexical records
    - 78 single words (base forms)
    - 23 multiwords (base forms)

    E0742227|chronic daily headache
    E0742228|cellobiose dehydrogenase
    E0742229|cervical disc herniation|cervical disk herniation
    E0742230|CDH
    E0742231|CDH|Cdh|cdh
    E0742232|cadherin1|cadherin 1|cadherin-1
    E0742233|CDH1|Cdh1|CDH-1
    E0742234|cadherin2|cadherin 2|cadherin-2
    ...
- Existing element words?

Existing Element Words

- N-grams
- Normalization and clustered
- Multiwords candidates
- Final validation

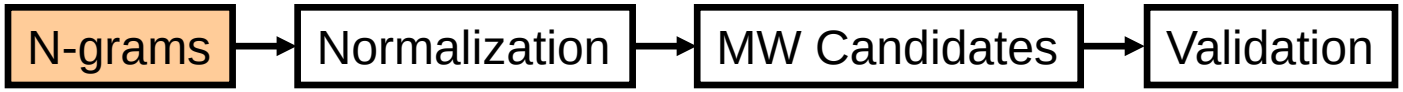
Existing Element Words



- N-grams:
  - TI and AB are tokenized into sentences
  - Sentences are then tokenized into n-gram words (space)
  - Outputs format:

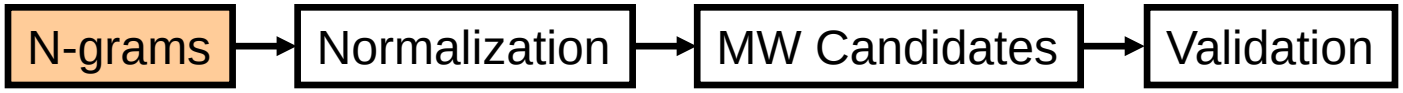
Document count	Word count	N-gram

Existing Element Words



- Range of N
- Filters: length and frequency of n-grams
- Issues: large scale (too big?)

Existing Element Words

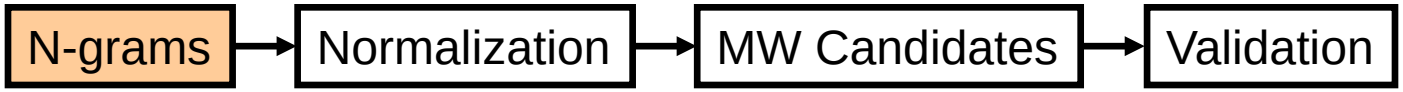


- Range of N (Lexicon.2014)

N	WC	Accu. 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%)
...	...	...

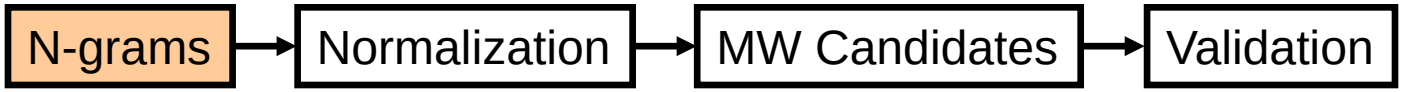
- Filters: length and frequency of n-grams
- Issues: large scale (too big?)

Existing Element Words



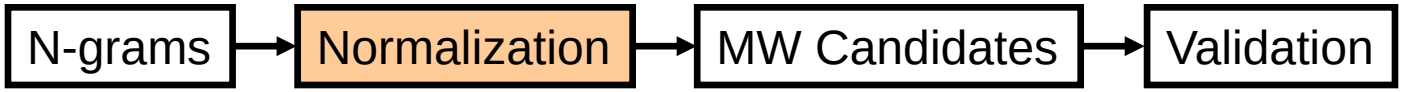
- Range of N: 1 ~ 5
- Filters: length and frequency of n-grams
  - length of N-grams: 50 (> 99.55%)
  - frequency: WC, DC, NWC
- Issues: large scale (too big?)

Existing Element Words



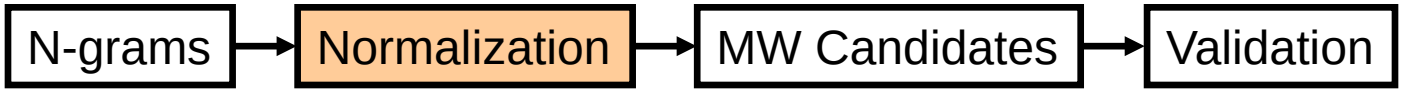
- Range of N: 1 ~ 5
- Filters: length and frequency of n-grams
- Issues: large scale (too big)
  - Increase memory size (hardware)
  - Persistence: DBM4, Tokyo cabinet, Java persistence APIs, etc.
  - Database: embedded or server
  - Algorithm: predictive filter, split & combine, modified data structure, etc.

Existing Element Words



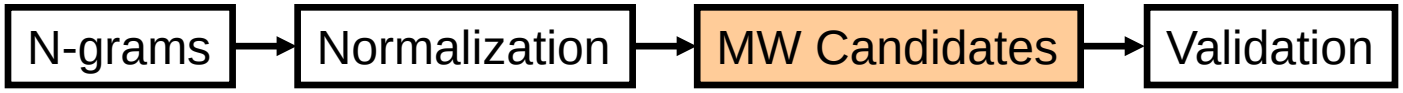
- A same term could be represented in many different forms
- Normalized by abstracting away from case (-f:l), genitive (-f:g), punctuation (-f:o) by Lexical Tools

Existing Element Words



- Example: diabetes mellitus
  - diabetes mellitus
  - diabetes mellitus,
  - diabetes mellitus]
  - diabetes mellitus:
  - diabetes mellitus.
  - [diabetes mellitus
  - diabetes mellitus)
  - (diabetes mellitus
  - (diabetes mellitus,
  - diabetes mellitus),
  - (diabetes mellitus;
  - diabetes mellitus?]
  - (diabetes mellitus)
  - diabetes mellitus -
  - Diabetes mellitus
  - Diabetes Mellitus
  - DIABETES MELLITUS
  - Diabetes mellitus,
  - Diabetes mellitus.
  - [Diabetes mellitus
  - [Diabetes Mellitus:
  - [Diabetes mellitus]
  - Diabetes Mellitus:
  - Diabetes Mellitus,
  - DIABETES MELLITUS]

Existing Element Words

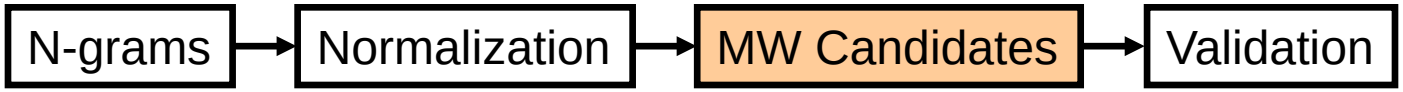


- Exclusive filters
- Inclusive filters
- Output: grouped by normalized forms for further analysis

Type	N	DC	WC	NWC	Norm N-Gram	N-Gram

N	DC	WC	NWC	Norm N-Gram	N-Gram

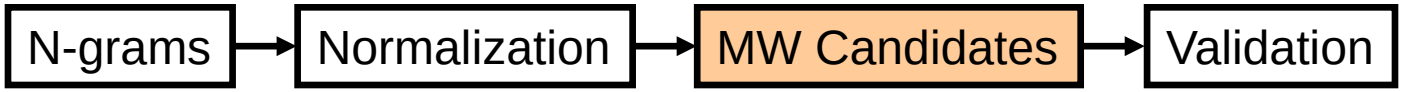
Existing Element Words



- Exclusive filters:

Filter	Descriptions	Examples
In Lexicon	(normalized) n-grams	[Diabetes mellitus]
Tail word pattern - Abbreviation	tail word – abbreviation/ acronym in parenthesis	mellitus (DM),
Tail word	preposition, conjunction, auxiliary, modal, determiner, complementizer	• mellitus in • diabetes mellitus or • mellitus is • diabetes mellitus may • mellitus and the • mellitus that
Head word	preposition, conjunction, auxiliary, determiner	• of diabetes mellitus • and diabetes mellitus: • have diabetes mellitus • that diabetes mellitus
Frequency	DC, WC, NWC	10, 50, 100
...	...	...

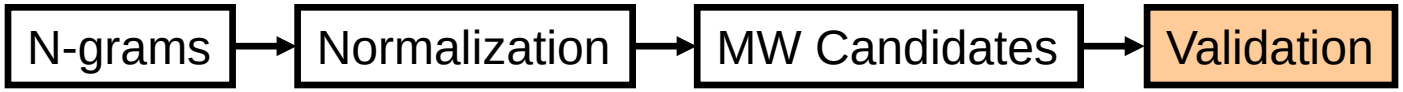
Existing Element Words



- Output example - mellitus:

Type	N	DC	WC	NWC	Norm N-Gram	N-Gram
…	…	…	…	…	…	…
LEX	2	**84089**	**120208**	**164764**	diabetes mellitus	diabetes mellitus
T_ABB	2	4692	4718	6693	diabetes mellitus dm	diabetes mellitus (DM)
T_CONJ	3	3164	3266	3500	diabetestmellitus and	diabetes mellitus, and
H_PREP	3	1991	2058	18912	of diabetes mellitus	of diabetes mellitus
…	…	…	…	…	…	…
TBD	4	661	885	1293	diabetes mellitus type 2	diabetes mellitus type 2
TBD	4	125	125	1293	diabetes mellitus type 2	diabetes mellitus type 2,
TBD	4	100	100	1293	diabetes mellitus type 2	diabetes mellitus type 2]
TBD	4	59	61	1293	diabetes mellitus type 2	Diabetes mellitus type 2
…	…	…	…	…	…	…

Existing Element Words

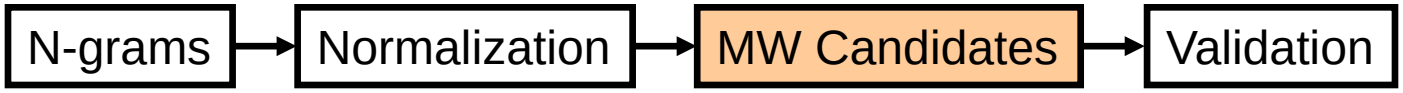


- Reviewed by linguists
- Add grammatical and lexical variant information for completed Lexicon records

Results – Existing Element Words

- Example: “mellitus”
  - Lexicon.2014: 24 lexical records
  - Candidate list:
    - 1304 n-grams contain “mellitus”
    - Retrieve 532 terms by exclusive filter (60% filter rate)
    - Mapped into 390 normalized forms (70% filter rate)
  - Results:
    - 36 new lexical records (150% increased)
      - 9 single words (base forms): mellitus, NODM, PNDM, etc.
      - 41 multiwords (base forms): diabetic mellitus, diabetes mellitus rat , new onset diabetes mellitus, new-onset diabetes mellitus, permanent neonatal diabetes mellitus, etc.
    - 7 associated existing records with 10 multiwords are updated for spelling variants and acronym expansion
      - noninsulin dependent diabetes mellitus|non-insulin dependent diabetes mellitus|non insulin-dependent diabetes mellitus
      - DM2| diabetes mellitus type 2

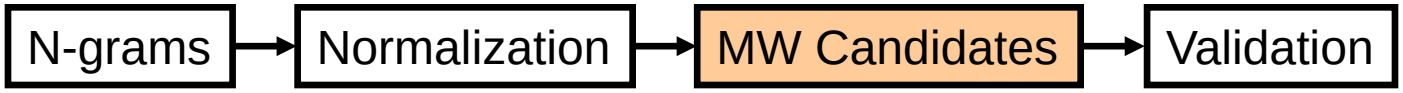
Existing Element Words



- **Inclusive filters** - spelling variants pattern:
  A term is most likely a valid word if it has spelling variant(s) in the N-grams

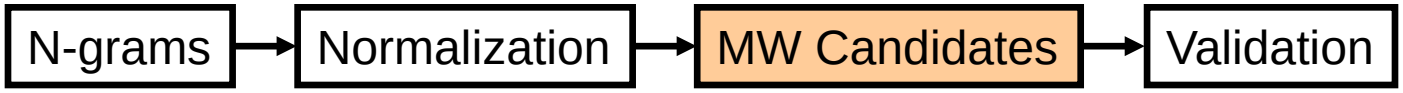
- Example:
  - o noninsulin dependent diabetes mellitus
  - o non-insulin dependent diabetes mellitus
  - o non insulin-dependent diabetes mellitus
  - o non insulin dependent diabetes mellitus
  - o noninsulin dependent diabetes mellitus

Existing Element Words



- Inclusive filters - Spelling variants pattern:
  - Spelling Variants Normalization
    - Non-ASCII: λmax|lambdamax
    - Synonym: St. Anthony's fire|Saint Anthony's fire
    - Spelling variant: CPA tumour|CPA tumor
    - Rank: Vth nerve|5th nerve
    - Number: 12-lead|twelve-lead
    - Roman Number: BoHV-I|BoHV-1
    - Punctuation: A.A.D.|AAD
    - Genitive: Laufe's forceps|Laufe forceps
    - Case: Latter-Day Saint|Latter-day Saint
    - Space: lattice work|latticework
  - MES (Metaphone, Edit distance, and Sorted distance)
  - ES (Edit distance and Sorted distance)

Existing Element Words

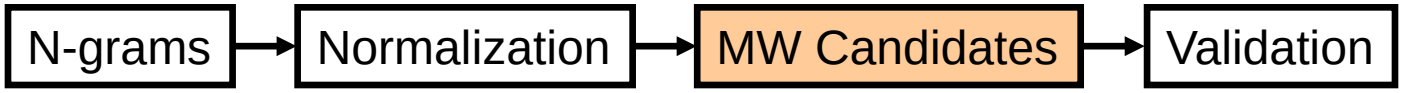


- Inclusive filters - Spelling variants pattern:
  - Spelling Variants Normalization
  - MES (Metaphone, Edit distance, and Sorted distance):
    Same Metaphone (LVG), limited edit distance, smallest sorted distance

Terms	Norm Terms	Metaphone	Edit Dist.
anemia|anaemia	anemia|anaemia	ANM	1
yuppie flu|yuppy flu	yuppie|yuppy	YPFL	2
directress|directrice	directress|directrice	TRKTRS	3
litchi nut|lychee nut	litchinut|lycheenut	LXNT	4

  - ES (Edit distance and Sorted distance)

Existing Element Words



- Inclusive filters - Spelling variants pattern:
  - Spelling Variants Normalization
  - MES (Metaphone, Edit distance, and Sorted distance):
  - ES (Edit distance and Sorted distance)

  Limited edit distance, smallest sorted distance

Terms	Norm Terms	Metaphone	Edit Dist.
ensoul|insoul	ensoul|insoul	ENSL|INSL	1
wholly|wholely	wholly|wholely	WL|WLL	1
racketball|racquetball	racketball|racquetball	RKTBL|RKKTBL	2
subtly|subtlely	subtly|subtlely	SBTL|SBTLL	2

Conclusion

- Established a systematic way to add more higher frequency words form MEDLINE to the SPECIALIST Lexicon

- Enhance the Lexicon's coverage on multiwords
  - Most high frequency single words (MEDLINE) are already included
  - Multiwords are an essential ingredient and play a key role in the success of NLP tasks

- Future works on multiwords

Questions

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