

# JAMIA Journal Club



# *JAMIA* Journal Club

- April 11<sup>th</sup>, 2019
- 3:00 – 4:00 PM, EDT
- Live webinar only available for CME credit
- See CME information on [amia.org](http://amia.org)



# How to Participate

- Speaker and Moderator converse for 35 minutes
- 25 minutes Q&A
- **Submit questions:**
  - Type into questions function box on lower right of screen within GoToWebinar
- Complete evaluation through emailed link



# Disclosures

The following speakers and planners, and their life partners, have no relevant relationships with commercial interests:

Speaker: Chris Lu

JAMIA: Michael Chiang, Kelson Zawack, Tiffany J Callahan

AMIA: Susanne Arnold, Pesha Rubinstein



# Learning Objective

After this live activity, the participant should be better able to:

- Understand different types of spell errors and corrections in consumer's questions
- Understand features of CSpell
- Better use and configure CSpell for NLP projects



# Co-manager: Kelson Zawack, PhD



Kelson Zawack is a Postdoc in the Biostatistics Department at Yale University. He is interested in inference of relationships between variables. He is currently working on methods for using fMRI to map the connections between brain regions that underlie reading and speaking.



# Co-manager and Moderator: Tiffany J. Callahan, MPH



Tiffany J. Callahan is a PhD candidate in the Computational Bioscience Program at the University of Colorado Denver Anschutz Medical Campus. Her research interests are biomedical knowledge representation and integration, high-throughput clinical phenotyping, and systems biology. Her work leverages artificial intelligence, statistical learning, and semantic web technologies

to deeply characterize and link pediatric lung clinical phenotypes and molecular endotypes.



# Q&A

- Type your Q into the questions function box on lower right of screen within GoToWebinar

***We encourage you to send questions during the presentation!***





# *JAMIA* Journal Club Selection

- Lu et al. Spell checker for consumer language (CSpell). J Am Med Inform Assoc. January 21, 2019  
– doi: 10.1093/jamia/ocy171
- **Speaker:**

**Chris J. Lu, PhD**

Lister Hill National Center for Biomedical  
Communications

National Library of Medicine

National Institutes of Health

Bethesda, MD



# Speaker: Chris J. Lu, PhD



**Dr. Chris J. Lu** is a senior systems architect in the Lexical Systems Group (LSG) at the NIH/NLM. Chris leads many natural language processing (NLP) projects at the NIH/NLM, including the public releases of the UMLS SPECIALIST Lexicon and Lexical Tools, Visual Tagging Tools (VTT), SubTerm Mapping Tools (STMT), MEDLINE N-gram Set, Text Categorization Tool (TC) and Spell Checker for Consumer Language (CSpell). Contact him at [chlu@mail.nih.gov](mailto:chlu@mail.nih.gov).



# Disclaimer

- The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.



# CSpell

- Spell checker for consumer language
- <http://umlslex.nlm.nih.gov/cSpell>
- <http://SPECIALIST.nlm.nih.gov/cSpell>
- Email: umlslex@nlm.nih.gov
- Questions (anytime)



# Acknowledgments

- Co-authors
  - Dr. Dina Demner-Fushman, Dr. Alan R. Aronson, Sonya E. Shooshan
- Colleagues
  - Dr. Halil Kilicoglu (Ensemble method), Dr. Russell Loane, Willie Rogers, James Mork, and Francois Lang
- Funding
  - NIH/NLM intramural research program



# Outline

- Introduction
  - Background & objectives
- Methodology
  - Training set – correction types
  - Correction techniques
  - Context-depend corrections
  - 2-stage ranking system
  - System design – multilayer
- Results, discussions, and conclusion



- Introduction
  - ✓ Background
  - ✓ Objectives
- Methodology
- Results and Discussion



# Background

- Health information consumers
  - Patients, families, caregivers, and the general public
  - Seek health information & ask questions online every day
- Sources of consumer health questions
  - MedlinePlus, forms and emails, etc.
  - Search engine, social media, forum, etc.
- Consumer questions
  - Contain many spelling errors, informal expressions, etc.
  - Spelling errors hinder automatic question answering
  - Spelling corrections are needed





# Consumer Questions Example

My mom is 82 years old suffering from **anixity** and depression for the last 10 years was **dianosed** early on set **deminita** 3 years ago. Do **yall** have a office in Greensboro NC? Can you recommend someone. she has **seretona** syndrome and **nonething** helps her. [2]

- Corrections:



- Reference:

[2] Kilicoglu H, Fiszman M, Roberts K, et al. An Ensemble method for spelling correction in consumer health questions. AMIA Annu Symp Proc., 2015: 727–36.

| Error     | Correction       |
|-----------|------------------|
| anixity   | <b>anxiety</b>   |
| dianosed  | diagnosed        |
| on set    | onset            |
| deminita  | <b>dementia</b>  |
| yall      | y'all            |
| seretona  | <b>serotonin</b> |
| nonething | nothing          |



# Ensemble Method (2015)

- Nonword

| Method   | Precision | Recall | F1   | $\delta$ F1 |
|----------|-----------|--------|------|-------------|
| ESpell   | 0.53      | 0.20   | 0.29 | -----       |
| Ensemble | 0.64      | 0.58   | 0.61 | <b>0.32</b> |

- Real-word

| Method   | Precision | Recall | F1   | $\delta$ F1 |
|----------|-----------|--------|------|-------------|
| ESpell   | 0.23      | 0.26   | 0.25 | -----       |
| Ensemble | 0.57      | 0.59   | 0.58 | <b>0.33</b> |



- Introduction
  - ✓ Background
  - ✓ Objectives
- Methodology
- Results and Discussion



# Project Objectives

- To develop a spelling tool to detect and correct all types of spelling errors in consumer language
  - Performance
  - Speed
  - Distributable
  - Generic
  - Open-source
  - Configurable



- Introduction
- Methodology
  - ✓ Training set – correction types
  - ✓ Correction techniques
  - ✓ Context-dependent corrections
  - ✓ 2-Stage ranking system
  - ✓ System design - multilayer
- Results and Discussion



# Types of Spelling Errors & Corrections

- Nonword vs. real-word errors
- Spelling errors
- Word boundary infraction errors: split and merge
- Punctuation errors
- ...
- Dictionary-based vs. non-dictionary-based corrections
- Isolated-word vs. context-dependent corrections
- Single token vs. multi-token corrections
- Others: informal expression, HTML/XML tag, ...
- ...
- Multiple corrections of combination errors



# Training Data Set

- Used both the training set and test set from Ensemble method for development

| Training Set Summary Statistics |        |
|---------------------------------|--------|
| Consumer health questions       | 471    |
| Tokens                          | 24,837 |
| Annotation tags                 | 1,008  |
| Nonword corrections             | 774    |
| Real-word corrections           | 964    |
| Word count/question             | 5-328  |
| Average word count/question     | 52.49  |
| Errors per/question             | 0-27   |
| Average error/question          | 2.14   |
| Error rate (error/token)        | 0.04   |



# Characteristics by Error Types

Distribution of errors in the training set

| Correction needed | Nonwords | Real words | ND    | Multiple <sup>a</sup> | Total by type |
|-------------------|----------|------------|-------|-----------------------|---------------|
| Spelling          | 348      | 153        | 113   | N/A                   | 614           |
| Merge             | 10       | 38         | 0     | N/A                   | 48            |
| Split             | 24       | 10         | 281   | N/A                   | 315           |
| Multiple          | N/A      | N/A        | N/A   | 31                    | 31            |
| Total             | 382      | 201        | 394   | 31                    | 1008          |
| Percentage        | 37.90    | 19.94      | 39.09 | 3.08                  | 100.00        |

Note: ND: not dictionary based.

<sup>a</sup>Errors that combine several types and require multiple corrections.





# Dictionary-based Examples

## Nonword (38%)

- Spelling:
  - **dianosed** => diag**n**osed
- Split:
  - **knowabout** => know about
- Merge:
  - **dur ing** => during

## Real-word (20%)

- Spelling:
  - **bowl** movement => bowel movement
- Split:
  - for **along** time => for a long time
- Merge:
  - diagnosed **on set** => diagnosed onset



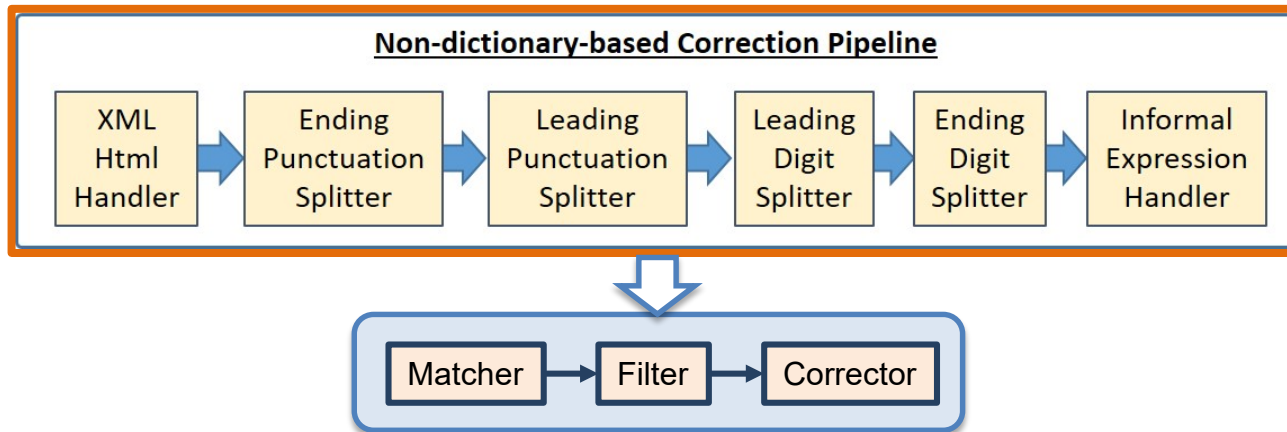
# Non-dictionary-based Examples

| Handler (11%)                                                                                                                                                                                                                                                                                                                                                        | Splitter (28%)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"><li>• XML/HTML:<ul style="list-style-type: none"><li>➤ <code>&amp;quot;germs&amp;quot;</code> =&gt; “germs”</li><li>➤ <code>&amp;amp;</code> =&gt; &amp;</li></ul></li><li>• Informal expression:<ul style="list-style-type: none"><li>➤ <code>pls</code> =&gt; please</li><li>➤ <code>whos</code> =&gt; who’s</li></ul></li></ul> | <ul style="list-style-type: none"><li>• Leading digit:<ul style="list-style-type: none"><li>➤ <code>1.5years</code> =&gt; 1.5 years</li><li>➤ <code>42nd</code></li></ul></li><li>• Ending digit:<ul style="list-style-type: none"><li>➤ <code>from2007</code> =&gt; from 2007</li><li>➤ <code>Co-Q10</code></li></ul></li><li>• Leading punctuation &amp;([[:<ul style="list-style-type: none"><li>➤ <code>volunteers(healthy)</code> =&gt; volunteers (healthy)</li><li>➤ <code>finger(s)</code></li></ul></li><li>• Ending punctuation : .?!,:;&amp;)]]:<ul style="list-style-type: none"><li>➤ <code>(..)why</code> =&gt; (..) why</li><li>➤ <code>NAD(P)H</code></li></ul></li></ul> |

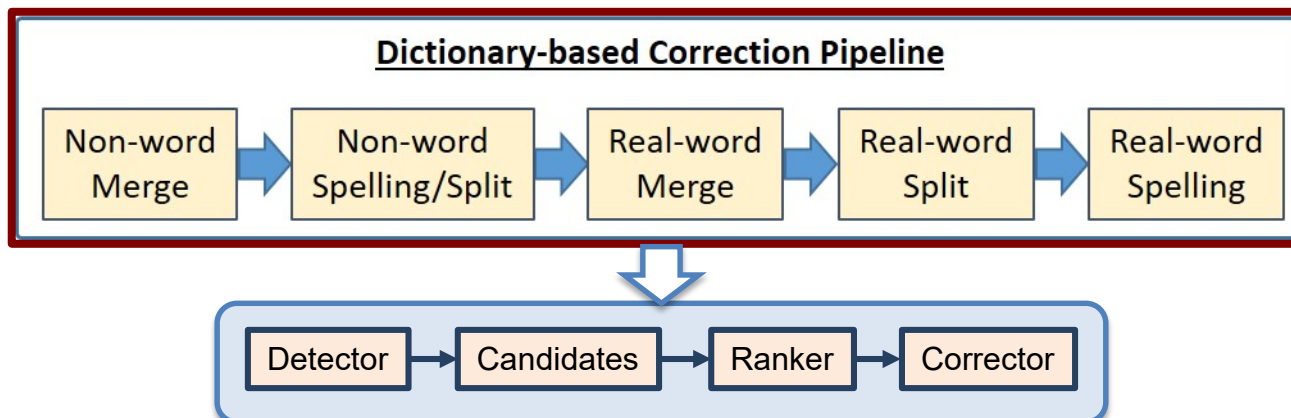


# Error & Correction Types

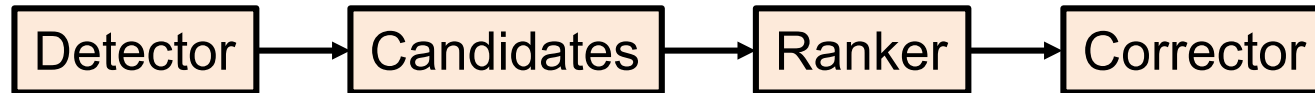
- Non-dictionary-based correction model:



- Dictionary-based correction model:



# Dictionary-based Model



- Example:

| Input    | Detector | Candidates                                                                                                                                                                    | Ranker                                                                                           | Corrector |
|----------|----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|-----------|
| diagnost | nonword  | <ul style="list-style-type: none"><li>• diagnose</li><li>• diagnosed</li><li>• diagnostic</li><li>• diagnosis</li><li>• diagnoses</li><li>• diagnoser</li><li>• ...</li></ul> | <ol style="list-style-type: none"><li>1) diagnosis</li><li>2) diagnosed</li><li>3) ...</li></ol> | diagnosis |



- Introduction
- Methodology
  - ✓ Training set – correction types
  - ✓ Correction techniques
  - ✓ Context-dependent corrections
  - ✓ 2-Stage ranking system
  - ✓ System design - multilayer
- Results and Discussion



# Edit Distance Similarity

- Edit **Distance**:  
The minimum number of operations required to transform one string into the other
- Example of “truly”:

| Input  | Output | Operation     | Dist. | Cost  | EDSS* |
|--------|--------|---------------|-------|-------|-------|
| truely | truly  | Deletion      | 1     | 0.096 | 0.904 |
| trly   | truly  | Insertion     | 1     | 0.090 | 0.910 |
| truli  | truly  | Substitution  | 1     | 0.100 | 0.900 |
| turly  | truly  | Transposition | 1     | 0.094 | 0.906 |

\* EDSS: Edit Distance Similarity Score



# Phonetic Similarity

- Phonetic algorithm:
  - Algorithm converts strings to codes for indexing by pronunciation
- Phonetic similarity:
  - Phonetic code + edit distance similarity
- Example: “diagnost” vs. “diagnosis”

| Phonetic Algorithm | diagnost   | diagnosis  | E.D. | PSS* |
|--------------------|------------|------------|------|------|
| Refined Soundex    | D6048036   | D60480303  | 2    | 0.80 |
| Metaphone          | TNST       | TNSS       | 1    | 0.90 |
| Caverphone 2       | TKNST11111 | TKNSS11111 | 1    | 0.90 |
| Double Metaphone   | TKNST      | TKNSS      | 1    | 0.90 |

\* PSS: Phonetic Similarity Score



# Overlap Similarity

- Overlap similarity:
  - Calculates the overlap of matching characters at the beginning and the end of 2 terms, divided by the length of the longer term
  - Similarity score is between 0.00 and 1.00
- Example: “truely” and “truly”

|                          | Values | Notes              |
|--------------------------|--------|--------------------|
| Lead Overlap Character   | 3      | tru-               |
| Trail Overlap Character  | 2      | -ly                |
| Max Length               | 6      | truley:6, truly: 5 |
| Overlap Similarity Score | 0.832  | $= (3+2)/6 = 5/6$  |





# Orthographic Similarity

- Orthographic (token) similarity
  - Looks and sounds alike (error model)
  - Enhanced similarity score (weighted sum)  
 $= 1.0 * \text{Edit Distance} + 0.7 * \text{Phonetic similarity} + 0.8 * \text{Overlap similarity}$
- Example: “tru<sup>e</sup>ly” and “truly”
  - Orthographic similarity score  
 $= 1.0 * 0.904 + 0.7 * 1.0 + 0.8 * 0.83$   
 $= 2.27$

|           | Edit Distance Similarity | Phonetic Similarity | Overlap Similarity |
|-----------|--------------------------|---------------------|--------------------|
| Operation | deletion                 | [TRL], [TRL]        | tru-, -ly          |
| Score     | $0.904 = 1 - 0.096$      | 1.0                 | $0.83 = (3+2)/6$   |
| Weights   | 1.0                      | 0.7                 | 0.8                |



# Word Frequency

- Word frequency score
  - Word count of occurrence within a given text corpus (language model)
  - Word: unigram (lowercase)
  - Normalized score: 0.0 ~ 1.0, normalized by the max. WC (the, 467,713 )
- Example:

| Word       | Word Count | Norm. Word Frequency |
|------------|------------|----------------------|
| diagnosis  | 9,083      | 0.019420029          |
| diagnosed  | 1,948      | 0.004164947          |
| diagnose   | 2,115      | 0.004522004          |
| diagnostic | 769        | 0.001644171          |
| diagnoses  | 203        | 0.000434027          |
| diagnoser  | 0          | 0.000000000          |



# Consumer Health Corpus

- Collected articles from 16 consumer-facing NIH websites
  - 17,139 articles
  - 10,228,699 tokens
  - 192,818 unique words

| Resources                                                             | Articles |
|-----------------------------------------------------------------------|----------|
| Genetic and Rare Diseases - diseases                                  | 6484     |
| Genetics Home Reference - conditions                                  | 1215     |
| Genetics Home Reference - genes                                       | 1439     |
| MedlinePlus - drugs                                                   | 1383     |
| MedlinePlus - medical encyclopedia                                    | 4425     |
| MedlinePlus - all health topics                                       | 1013     |
| MedlinePlus - herbs and supplements                                   | 153      |
| National Eye Institute                                                | 36       |
| National Heart, Lung, and Blood Institute                             | 141      |
| National Institute of Allergy and Infectious Diseases                 | 53       |
| National Institute of Arthritis and Musculoskeletal and Skin Diseases | 55       |
| National Institute of Child Health and Human Development              | 81       |
| National Institute on Deafness and Other Communication Disorders      | 15       |
| National Institute of Diabetes and Digestive and Kidney Disease       | 181      |
| National Institute of Mental Health                                   | 26       |
| National Institute of Neurological Disorders and Stroke               | 439      |



- Introduction
- Methodology
  - ✓ Training set – correction types
  - ✓ Correction techniques
  - ✓ Context-dependent corrections
  - ✓ 2-Stage ranking system
  - ✓ System design - multilayer
- Results and Discussion



# Toward Context-dependent Corrections

- Isolated-word corrections

| Input    | Technique      | Correction |
|----------|----------------|------------|
| diagnost | Orthographic   | diagnose   |
| diagnost | Word frequency | diagnosis  |
| diagnost | Noisy Channel  | diagnosis  |

- Context-dependent corrections

| Input        | Correction    |
|--------------|---------------|
| the diagnost | the diagnosis |
| was diagnost | was diagnosed |



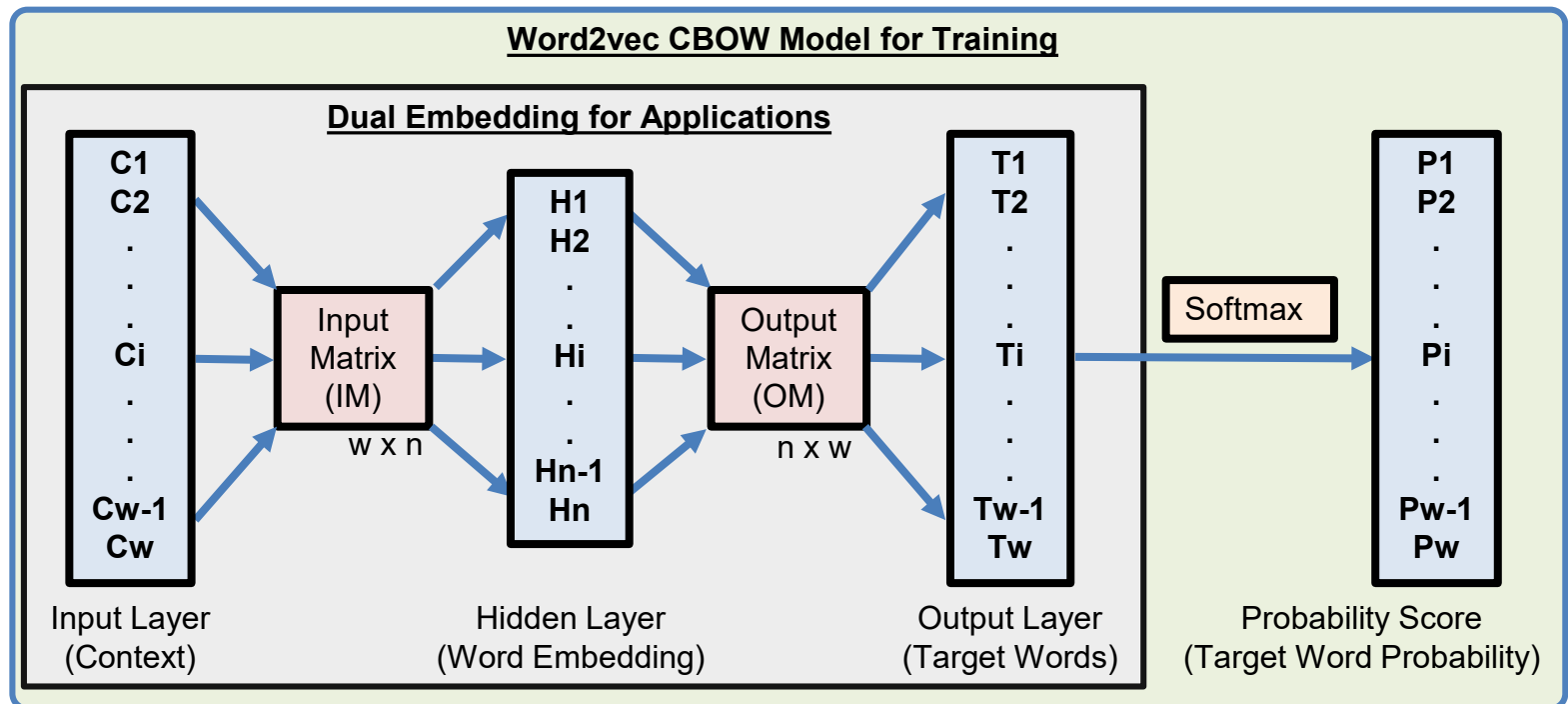
# Word2vec: Word Embedding

- Word vectors (word2vec, 2013 <sup>[22-23]</sup>):
  - Each word has an associated vector
  - Represent the meaning of a word in some abstract way
  - Capture meaningful syntactic and semantic regularities
- 2 Models:
  - Continuous bag-of-words (CBOW): predict a word from context
  - Continuous skip-gram: predict context from a word
- Examples (prediction by similarity):
  - Man -> Kings, Woman -> **Queens**
  - =>  $\text{Kings} - \text{Man} + \text{Woman} = \text{Queens}$  (vector operations)



# CBOW - Dual Embedding

- Dual embedding: use both the IM and OM matrices to compute context scores of the predicted target word with given contexts
- Context score =  $[T] = [C] \times [IM] \times [OM]$
- Improve performance (9.13% in F1)



# Single vs. Dual Embedding

- Single embedding:
  - Use the [IM] as word vectors
  - Use the cosine similarity between context and candidates' word vectors for context score <sup>[2,15]</sup>
- Dual embedding:
  - Use both [IM] and [OM] to feed the context into the original CBOW model for context score of candidates

| Embedding | Matrixes | Precision | Recall | F1 *          |
|-----------|----------|-----------|--------|---------------|
| Single    | IM       | 0.5887    | 0.5917 | 0.5902        |
| Dual      | IM & OM  | 0.8035    | 0.5917 | <b>0.6815</b> |

Dual embedding has better precision and F1 (+9.13%)

\* Training set, nonword test by context scores





# Context-dependent Corrections

- Nonword corrections

| Input               | Correction           |
|---------------------|----------------------|
| the <b>diagnost</b> | the <b>diagnosis</b> |
| was <b>diagnost</b> | was <b>diagnosed</b> |

- Real-word corrections

| Input Text             | Correction             |
|------------------------|------------------------|
| <b>smell</b> size      | <b>small</b> size      |
| <b>smell</b> amount    | <b>small</b> amount    |
| <b>smell</b> intestine | <b>small</b> intestine |

| Input Text             | Correction             |
|------------------------|------------------------|
| foul <b>small</b>      | foul <b>smell</b>      |
| <b>small</b> an order  | <b>smell</b> an order  |
| taste and <b>small</b> | taste and <b>smell</b> |



- Introduction
- Methodology
  - ✓ Training set – correction types
  - ✓ Correction techniques
  - ✓ Context-dependent corrections
  - ✓ 2-Stage ranking system
  - ✓ System design - multilayer
- Results and Discussion



# 2-stage Ranking System

- Efficiently utilize the knowledge sources, similar to regular season/play-offs in sports championship
- Stage-1 (regular season):
  - Orthographic similarity scores  $>$  threshold
- Stage-2 (playoffs):
  - Chain comparators by the context score, then the noisy channel score in a sequential order
  - Rank in stage-1 is ignored



# 1-stage vs. 2-stage Ranking

- 2-stage is better than 1-stage
  - Nonword, training set

| Stage-1        | Stage-2                      | Precision     | Recall        | F1 *          |
|----------------|------------------------------|---------------|---------------|---------------|
| Orthographic   | N/A                          | 0.7606        | 0.7636        | <b>0.7621</b> |
| Word Frequency | N/A                          | 0.6970        | 0.6925        | 0.6948        |
| Noisy Channel  | N/A                          | 0.7134        | <b>0.7171</b> | 0.7152        |
| Context Score  | N/A                          | <b>0.8035</b> | 0.5917        | 0.6815        |
| Ensemble       | N/A                          | 0.7516        | 0.7545        | 0.7531        |
| Orthographic   | Word Frequency               | 0.8241        | 0.7687        | 0.7955        |
| Orthographic   | Noisy Channel                | 0.8255        | <b>0.7700</b> | 0.7968        |
| Orthographic   | Context Score                | <b>0.8996</b> | 0.5672        | 0.6957        |
| Orthographic   | Context Score, Noisy Channel | 0.8047        | 0.7842        | <b>0.8115</b> |

\* Training set, nonword test

Best F1



# 2-stage Ranking Example

- Input: havy
- Candidates: 441
  - Stage 1: find qualified candidates by orthographic similarity score
    - heavy (2.25), have (2.20), hay (2.13), and wavy (2.13), ...
    - Threshold (92%):  $2.25 \times 0.92 = 2.07$
    - happy (1.92), hair (1.83), and lady (1.56), ...
  - Stage 2: use context score, then Noisy Channel score

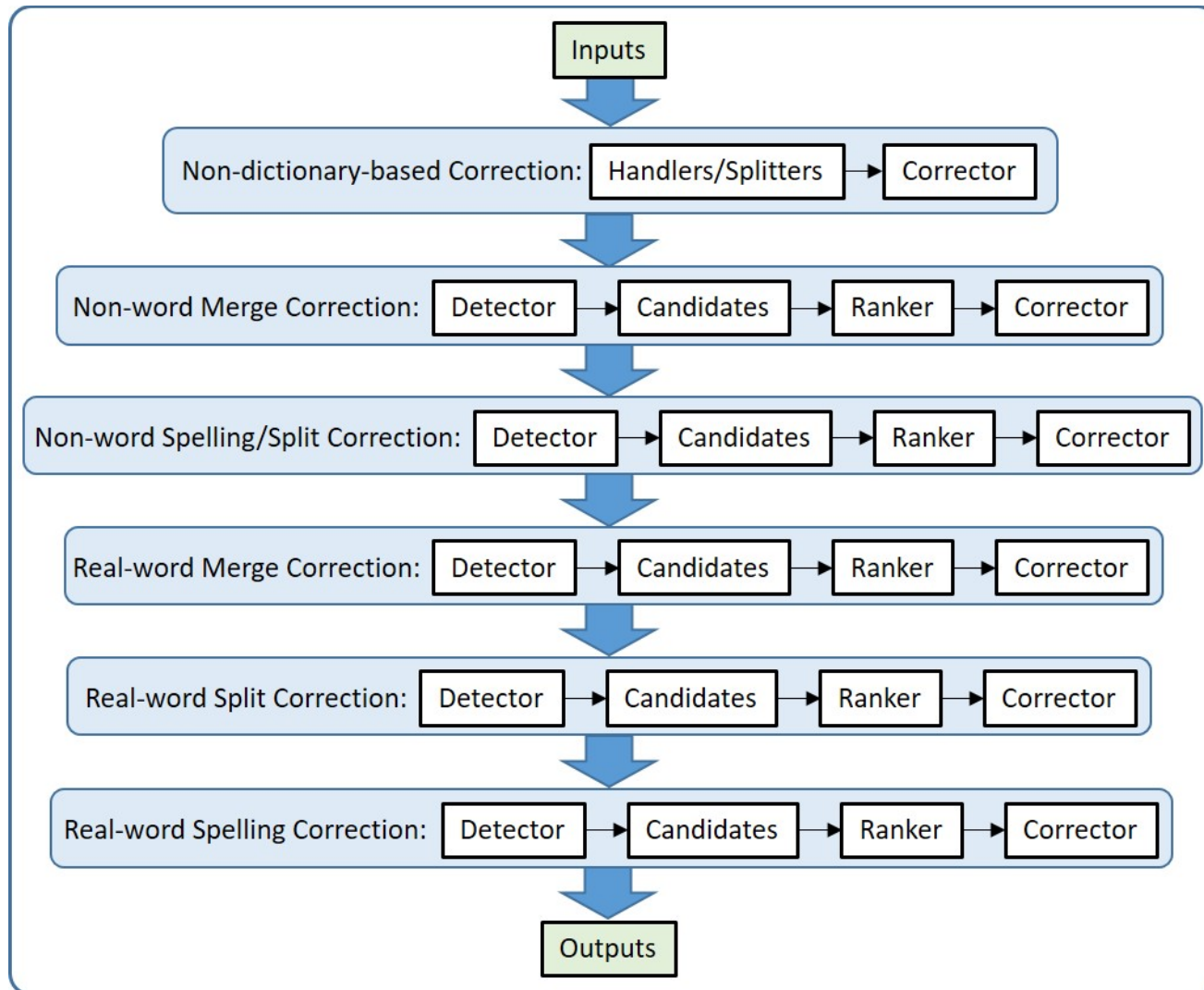
| Input Text    | Correction    | Token Scores | Context Scores |         |         |         | N.C. Scores |
|---------------|---------------|--------------|----------------|---------|---------|---------|-------------|
|               |               |              | heavy          | have    | hay     | wavy    |             |
| havy          | have*         | 2.20         | 0.0000         | 0.0000  | 0.0000  | 0.0000  |             |
| havy duty     | heavy duty    | 2.25         | 0.0597         | -0.0302 | -0.0053 | 0.0074  | 0.00198     |
| havy diabetes | have diabetes | 2.20         | -0.0067        | 0.0586  | -0.0518 | -0.0813 | 0.14933     |
| havy fever    | hay fever     | 2.13         | -0.1331        | 0.2280  | 0.2292  | -0.0391 | 0.00032     |
| havy lines    | wavy lines    | 2.13         | -0.0170        | -0.0410 | -0.0702 | 0.1495  | 0.00004     |



- Introduction
- Methodology
  - ✓ Training set – correction types
  - ✓ Correction techniques
  - ✓ Context-dependent corrections
  - ✓ 2-Stage ranking system
  - ✓ System design - multilayer
- Results and Discussion



# Multilayer Design



# CSpell Algorithm Summary-1

| Process                                                        | Detector                                                                                                                                                                                                                                           | Candidates                                                                                                                                                                                                                                              | Ranker                                                                                                                                                                                                                       | Corrector                                                                                                    | Features                                                                                                                                                                                                                                                |
|----------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1. <u>Non-Dic</u> ,<br><u>PreCorrector</u><br>(Java 8 Stream)  | <ul style="list-style-type: none"> <li>Pattern Match</li> <li>HashMap</li> </ul>                                                                                                                                                                   | <ul style="list-style-type: none"> <li>Algorithm</li> <li>HashMap</li> </ul>                                                                                                                                                                            | <ul style="list-style-type: none"> <li>None</li> </ul>                                                                                                                                                                       | <ul style="list-style-type: none"> <li>Replace</li> <li><u>FlatMap</u></li> <li><u>UpdateHist</u></li> </ul> | <ul style="list-style-type: none"> <li><u>XmlHtmlHandler</u></li> <li><u>EndPuncSplitter</u></li> <li><u>LeadingPuncSplitter</u></li> <li><u>LeadingDigitSplitter</u></li> <li><u>EndingDigitSplitter</u></li> <li><u>InformalExpHandler</u></li> </ul> |
| 2. <u>Dic</u> , <u>Non-Word</u> ,<br><u>Merge</u>              | <ul style="list-style-type: none"> <li>Not Valid Word: <u>splitDic (SwNoAaLc.dic)</u></li> <li>Not Exceptions: <u>Digit</u>, <u>Punc</u>, <u>Url</u>, <u>Email</u>, <u>EmptyStr</u>, <u>UpperCase</u>, <u>1Char</u>, <u>Measurement</u></li> </ul> | <ul style="list-style-type: none"> <li><math>\leq</math> <u>MergeNo</u></li> <li>Merge " " and "-"</li> <li>context not an exception</li> <li><u>orgWord</u> not in <u>mwDic</u></li> <li><u>cand</u> in <u>suggDic</u>, not in <u>aaDic</u></li> </ul> | <ul style="list-style-type: none"> <li><u>Word2Vec: topScore</u> <math>\neq</math> 0</li> <li>Frequency</li> </ul>                                                                                                           | <ul style="list-style-type: none"> <li>Reconstruct the whole text</li> <li><u>UpdateHist</u></li> </ul>      |                                                                                                                                                                                                                                                         |
| 3. <u>Dic</u> , <u>Non-Word</u> ,<br><u>1-to-1 &amp; Split</u> | <ul style="list-style-type: none"> <li>Not Valid Word: <u>CheckDic (Ew, Num)</u></li> <li>Not Exceptions: <u>Digit</u>, <u>Punc</u>, <u>Url</u>, <u>Email</u>, <u>EmptyStr</u>, <u>Measurement</u></li> </ul>                                      | <ul style="list-style-type: none"> <li><u>Dist</u> <math>\leq</math> 2</li> <li>In <u>suggDic</u></li> </ul>                                                                                                                                            | <ul style="list-style-type: none"> <li>Sorted by Orthographic</li> <li>Refined candidates to a range (within 0.08 of <u>top Oscore</u>)</li> <li>Use top context score</li> <li>Use top <u>NoisyChannel</u> score</li> </ul> | <ul style="list-style-type: none"> <li>Replace</li> <li><u>FlatMap</u></li> <li><u>UpdateHist</u></li> </ul> |                                                                                                                                                                                                                                                         |





# CSpell Algorithm Summary-2

| Process                   | Detector                                                                                                                                                                                                                                            | Candidates                                                                                                                                                                                                                                                                                                                                                                                                                                            | Ranker                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Corrector                                                                                                                    |
|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|
| 4. Dic, Real-Word, Merge  | <ul style="list-style-type: none"> <li>Not corrected yet</li> <li>Valid Word (<u>splitDic</u>)</li> <li>Not Exceptions</li> </ul>                                                                                                                   | <ul style="list-style-type: none"> <li><math>\leq</math> MergeNo</li> <li>Merge “ ”, not “-”</li> <li>context not an exception</li> <li>orgWord not in mwDic,</li> <li>cand in suggDic, not in aaDic</li> <li>Cand has word2Vec</li> <li>Cand WC <math>\geq</math> 10 (Conf)</li> <li>No short word merge</li> </ul>                                                                                                                                  | <ul style="list-style-type: none"> <li>Use Word2Vec to find top rank:</li> <li>Valid top rank: <ul style="list-style-type: none"> <li>- <math>\text{orgS} &gt; 0 \ \&amp; \ \text{topS} * \text{Fac} &gt; \text{orgS}</math></li> <li>- <math>\text{orgS} &lt; 0 \ \&amp; \ \text{topS} &gt; 0</math></li> <li>- <math>\text{orgS} &lt; 0 \ \&amp; \ \text{topS} &lt; 0 \ \&amp; \ \text{topS} &gt; \text{orgS} * \text{Fac}</math></li> </ul> </li> <li>cFac = 0.60 (Conf)</li> </ul>                                                                                                                                                                                                                                                                                          | <ul style="list-style-type: none"> <li>Reconstruct the whole text</li> <li>UpdateHist</li> </ul>                             |
| 5. Dic, Real-Word, Split  | <ul style="list-style-type: none"> <li>Not corrected yet</li> <li>Valid Word (<u>checkDic</u>)</li> <li>Not Exceptions</li> <li>Length <math>&gt; 3</math> (Conf)</li> <li>Has word2Vec</li> <li>WC <math>&gt; 200</math> (Conf)</li> </ul>         | <ul style="list-style-type: none"> <li><math>\leq</math> SplitNo</li> <li>The split term is LMW</li> <li>Short split word <ul style="list-style-type: none"> <li>- total no of short split word</li> </ul> </li> <li>split word <ul style="list-style-type: none"> <li>- Valid split words (<u>splitDic</u>)</li> <li>- Has word2Vec</li> <li>- WC <math>\geq 200</math> (Conf)</li> <li>- Not unit</li> <li>- Not proper noun</li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>Use Word2Vec to find top rank:</li> <li>Valid top rank: <ul style="list-style-type: none"> <li>- Same score ranking rules as merge</li> </ul> </li> <li>cFac = 0.00 (Conf)</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | <ul style="list-style-type: none"> <li>Add the FlatMap of the split term to the in token list</li> <li>UpdateHist</li> </ul> |
| 6. Dic, Real-Word, 1-to-1 | <ul style="list-style-type: none"> <li>Not corrected yet</li> <li>Valid Word (<u>checkDic</u>)</li> <li>Not Exceptions (aa, pn)</li> <li>Length <math>\geq 2</math> (Conf)</li> <li>Has word2Vec</li> <li>WC <math>&gt; 65</math> (Conf)</li> </ul> | <ul style="list-style-type: none"> <li>Valid word (<u>suggDic</u>)</li> <li>Cand has word2Vec</li> <li>Cand WC <math>\geq 0</math> (Conf)</li> <li>Cand Length <math>\geq 2</math> (Conf)</li> <li>Not inflVar of inWord</li> <li>Heuristics Rules <ul style="list-style-type: none"> <li>- pmDist=0 &amp; prDist=0</li> <li>- tDist1 <math>&lt; 3</math> &amp; tDist2 <math>&lt; 4</math> &amp; pmDist == 0</li> </ul> </li> </ul>                   | <ul style="list-style-type: none"> <li>Find top rank cand (sorted): <ul style="list-style-type: none"> <li>- top Orthographic Score</li> <li>- top freqScore</li> <li>- top EditScore</li> <li>- top phoneticScore</li> <li>- top overlapScore</li> </ul> </li> <li>Validate top rank cand: <ul style="list-style-type: none"> <li>- cScore (top <math>&gt; 0</math>, org <math>&lt; 0</math>) <ul style="list-style-type: none"> <li>* top <math>&gt; -0.1 * \text{org}</math></li> <li>* top-org <math>&gt; 0.085</math></li> <li>* org <math>&gt; -0.085</math></li> <li>* topF <math>&gt; 0.0006</math></li> <li>* topF/orgF <math>&gt; 0.035</math></li> <li>* orgF - topF <math>&lt; 0.02</math> or topF <math>&gt; \text{orgF}</math></li> </ul> </li> </ul> </li> </ul> |                                                                                                                              |



# CSpell Correction Examples

| Ex-1: Input Text                                                | Output: different types of corrections                                                |
|-----------------------------------------------------------------|---------------------------------------------------------------------------------------|
| He was <b>dianosed</b> early <b>on set deminita 3years</b> ago. | He was <u>diagnosed</u> early <u>onset</u> <u>dementia</u> <u>3 years</u> ago.        |
|                                                                 | <div>NW Spelling</div> <div>RW Merge</div> <div>NW Spelling</div> <div>ND Split</div> |

| Ex-2: Input Text                | Output: multiple corrections            |
|---------------------------------|-----------------------------------------|
| I have a <b>shuntfrom2007</b> . | I have a <u>shunt from 2007</u> .       |
|                                 | <div>NW Split</div> <div>ND Split</div> |

| Ex-3: Input Text                                                                | Output: multiple corrections                                                    |
|---------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| I am permanently depressed and was on 2 or 3 different <b>anti depresants</b> . | I am permanently depressed and was on 2 or 3 different <u>antidepressants</u> . |
|                                                                                 | <div>RW Merge</div> <div>NW Spelling</div>                                      |



- Introduction
- Methodology
- Results and Discussion
  - ✓ Training set
  - ✓ Test sets
  - ✓ Discussion and conclusions



# Results From Training Set

- Detection:

| Method   | Precision | Recall | F1            | $\delta$ F1 |             |             |
|----------|-----------|--------|---------------|-------------|-------------|-------------|
| ESpell   | 0.3475    | 0.4253 | 0.3825        | -----       | -0.11       | -0.31       |
| Jazzy    | 0.8499    | 0.3465 | 0.4923        | 0.11        | -----       | -0.20       |
| Ensemble | 0.8078    | 0.6017 | 0.6897        | 0.31        | 0.20        | -----       |
| CSpell   | 0.9289    | 0.7178 | <b>0.8098</b> | <b>0.42</b> | <b>0.32</b> | <b>0.12</b> |

- Correction:

| Method   | Precision | Recall | F1            | $\delta$ F1 |             |             |
|----------|-----------|--------|---------------|-------------|-------------|-------------|
| ESpell   | 0.2076    | 0.2541 | 0.2285        | -----       | -0.05       | -0.38       |
| Jazzy    | 0.4860    | 0.1981 | 0.2815        | 0.05        | -----       | -0.33       |
| Ensemble | 0.7201    | 0.5363 | <b>0.6147</b> | 0.38        | <b>0.33</b> | -----       |
| CSpell   | 0.8416    | 0.6504 | <b>0.7338</b> | <b>0.50</b> | <b>0.45</b> | <b>0.12</b> |



- Introduction
- Methodology
- Results and Discussion
  - ✓ Training set
  - ✓ Test sets
  - ✓ Discussion and conclusions



# Test set

- Consumer health questions with most OOV (out of vocabulary) terms
  - 2 annotators, 1 arbitrator
  - Discrepancies: reconciled, then arbitrated

|                             | Test          | Training |
|-----------------------------|---------------|----------|
| Consumer health questions   | <b>224</b>    | 471      |
| Tokens                      | <b>16,707</b> | 24,837   |
| Annotation tags             | 1,946         | 1,008    |
| Nonword corrections         | 974           | 774      |
| Real-word corrections       | 1178          | 964      |
| Word count/question         | 3-337         | 5-328    |
| Average word count/question | 72.36         | 52.49    |
| Errors/question             | 0-22          | 0-27     |
| Average error/question      | <b>4.90</b>   | 2.14     |
| Error rate (error/token)    | <b>0.07</b>   | 0.04     |



# Results From Test Set

- Detection:

| Method   | Precision | Recall | F1     | $\delta$ F1 |
|----------|-----------|--------|--------|-------------|
| Ensemble | 0.8210    | 0.5645 | 0.6690 | -----       |
| CSpell   | 0.8900    | 0.7419 | 0.8093 | <b>0.14</b> |

- Correction:

| Method   | Precision | Recall | F1     | $\delta$ F1 |
|----------|-----------|--------|--------|-------------|
| Ensemble | 0.6975    | 0.4796 | 0.5684 | -----       |
| CSpell   | 0.7607    | 0.6341 | 0.6917 | <b>0.12</b> |



# Other Tests

- Important terms on test set
  - Important terms for question understanding
  - Identified during manual annotation
  - Real-word, F1 = 0.6739
- MEDLINE titles
  - 2,585 MEDLINE titles with errors
  - F1 = 0.75 (nonword)
  - Used for other biomedical spelling error corrections
- Consumer Health Information Question Answering (CHIQA)
  - Integrated in CHIQA
  - Many spelling errors were found and corrected





# Speed Performance Test

- Real-time corrections
  - 11.38 faster than Ensemble (= 2064/181.28)
  - 430 words with 12.6 nonword corrections per second
  - 120 words with 3.6 real-word corrections per second
  - Tested on Red Hat Enterprise Workstation 7.3 Maipo (Red Hat Inc, Raleigh, NC), Intel Xeon(R) CPU E5506 @2.13GHz (Dell)



- Introduction
- Methodology
- Results and Discussion
  - ✓ Training set
  - ✓ Test sets
  - ✓ Discussion and conclusions



# Dictionary Selection

- Relevant, supervised and comprehensive dictionary is better

| Dictionary | Size    | B*  | S*  | C*  | Precision | Recall | F1**   |
|------------|---------|-----|-----|-----|-----------|--------|--------|
| Jazzy      | 159,345 | No  | No  | No  | 0.2085    | 0.6835 | 0.3195 |
| Ensemble   | 459,038 | Yes | No  | No  | 0.7139    | 0.7610 | 0.7367 |
| MEDLINE    | 496,387 | Yes | No  | No  | 0.7506    | 0.7468 | 0.7487 |
| Lexicon    | 558,353 | Yes | Yes | Yes | 0.8407    | 0.7842 | 0.8115 |

\*

- B: Biomedical terms
- S: Supervised (manually verified)
- C: Comprehensive lexical information (abbreviations, acronyms, proper nouns, multiwords, etc.)

\*\* Training set, nonword test



# Corpus Selection

- Relevant (not size) corpus is better

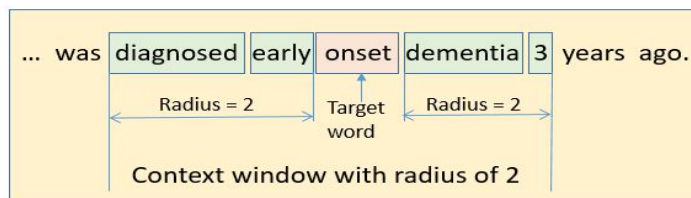
|                  | Consumer Health Corpus  | Medline N-gram Set |
|------------------|-------------------------|--------------------|
| Resources        | 16 NIH public web sites | MEDLINE            |
| Articles         | 17,139                  | 26,759,399         |
| Sentences        | 550,193                 | 163,021,640        |
| Tokens           | 10,228,699              | 3,386,661,350      |
| Unique Word (LC) | 109,818                 | 496,388            |
| Dictionary words | 8.5886%                 | 37.8507%           |
| Precision        | 0.8407                  | 0.8085             |
| Recall           | 0.7842                  | 0.7907             |
| F1*              | 0.8115                  | 0.7995             |

\* Training set, nonword test



# Context Window Size Selection

- Training model (CBOW):
  - window size = 5
- Application:
  - context radius = 2



- Local context is more important than global context in spelling correction application

| Context Radius | Precision | Recall | F1            |
|----------------|-----------|--------|---------------|
| 1              | 0.8380    | 0.7817 | 0.8088        |
| 2              | 0.8407    | 0.7842 | <b>0.8115</b> |
| 3              | 0.8366    | 0.7804 | 0.8075        |
| 4              | 0.8352    | 0.7791 | 0.8061        |
| 5              | 0.8352    | 0.7791 | 0.8061        |
| 6              | 0.8296    | 0.7739 | 0.8008        |
| ...            | ...       | ...    | ...           |
| 10             | 0.8296    | 0.7739 | 0.8008        |
| 25             | 0.8283    | 0.7726 | 0.7995        |
| 50             | 0.8283    | 0.7726 | 0.7995        |
| 100            | 0.8283    | 0.7726 | 0.7995        |



# Contribution Summary

- A deployable context-dependent correction tool
  - Dual embedding of CBOW model
- Correct all types of errors in consumer language
  - A multilayer design of spelling corrections
- Effectively utilize knowledge sources
  - 2-stage ranking system



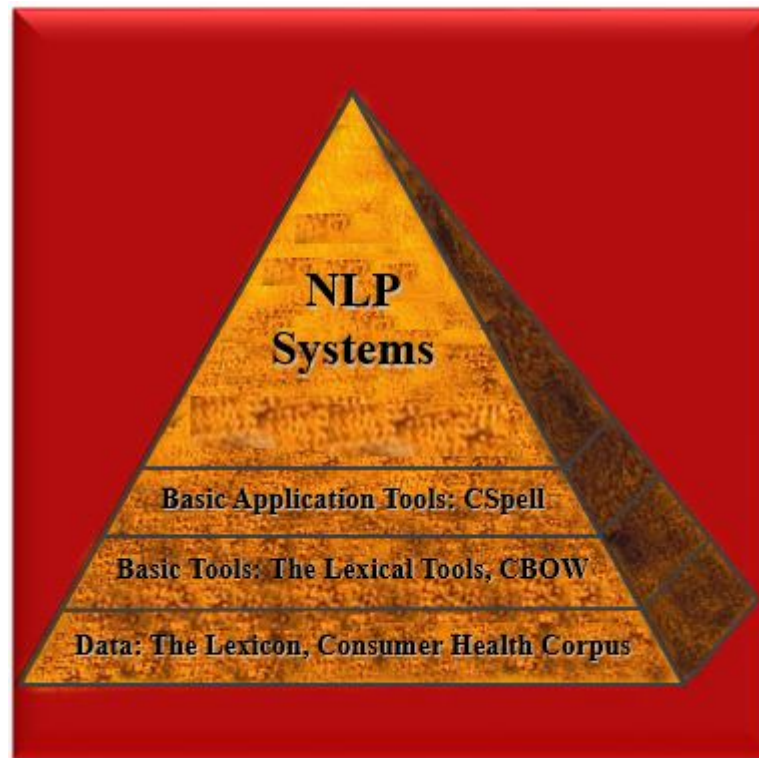
# CSpell

- Provides many configurable options
  - Dictionary
  - Corpus
  - Types of corrections
- Configuration file
  - Default with empirical best values of thresholds and other variables
- Command line tool and Java APIs



# Conclusion

- Provide an generic, stand-alone, configurable, open-source spell checker for NLP community





# CSpell Questions

- Spell checker for consumer language
  - <http://umlslex.nlm.nih.gov/cSpell>
  - <http://SPECIALIST.nlm.nih.gov/cSpell>
  - <http://umlslex.nlm.nih.gov>



# Q&A

- Type your Q into the questions function box on lower right of screen
- ...
- At end of webinar, please complete evaluation in chat box area



# CME Certificate

- Physicians: *For live webinar only*, claim your credit via the evaluation survey; receive pdf of your CME certificate via email



# Upcoming *JAMIA*-JC

- The next *JAMIA*-JC will be on **Thursday, May 9<sup>th</sup>, 2019, 3:00 – 4:00 EDT**
- Yuan C, Ryan PB, Ta C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc*. 2019 Apr 1;26(4):294-305.



This *JAMIA* Journal Club has now concluded.

[www.amia.org](http://www.amia.org)

