



Spelling Correction in Clinical Notes

Jon Patrick
Mojtaba Sabbagh
Suvir Jain
Haifeng Zheng



Clinical Notes

- Collections
 - 6 years from ICU - 60 M tokens
 - 4 years from ED - 20M tokens
 - Other smaller sets



NLP Tasks

- High Accuracy IR
- High Accuracy IE
- High Accuracy QA
- In Intensive Care Units



Clinical Notes Problems

- 30% non-words
- Poor spelling
- Poor grammar
- Personals; spellings, acronyms, abbrevs
- Clinical terminology – highly productive
- Measurements and Scores: 3-5+mcml/h



Spelling Correction

- The large medical language makes traditional correctors inadequate
- Need spelling correction as a part of our process of Knowledge Discovery and Knowledge Reuse

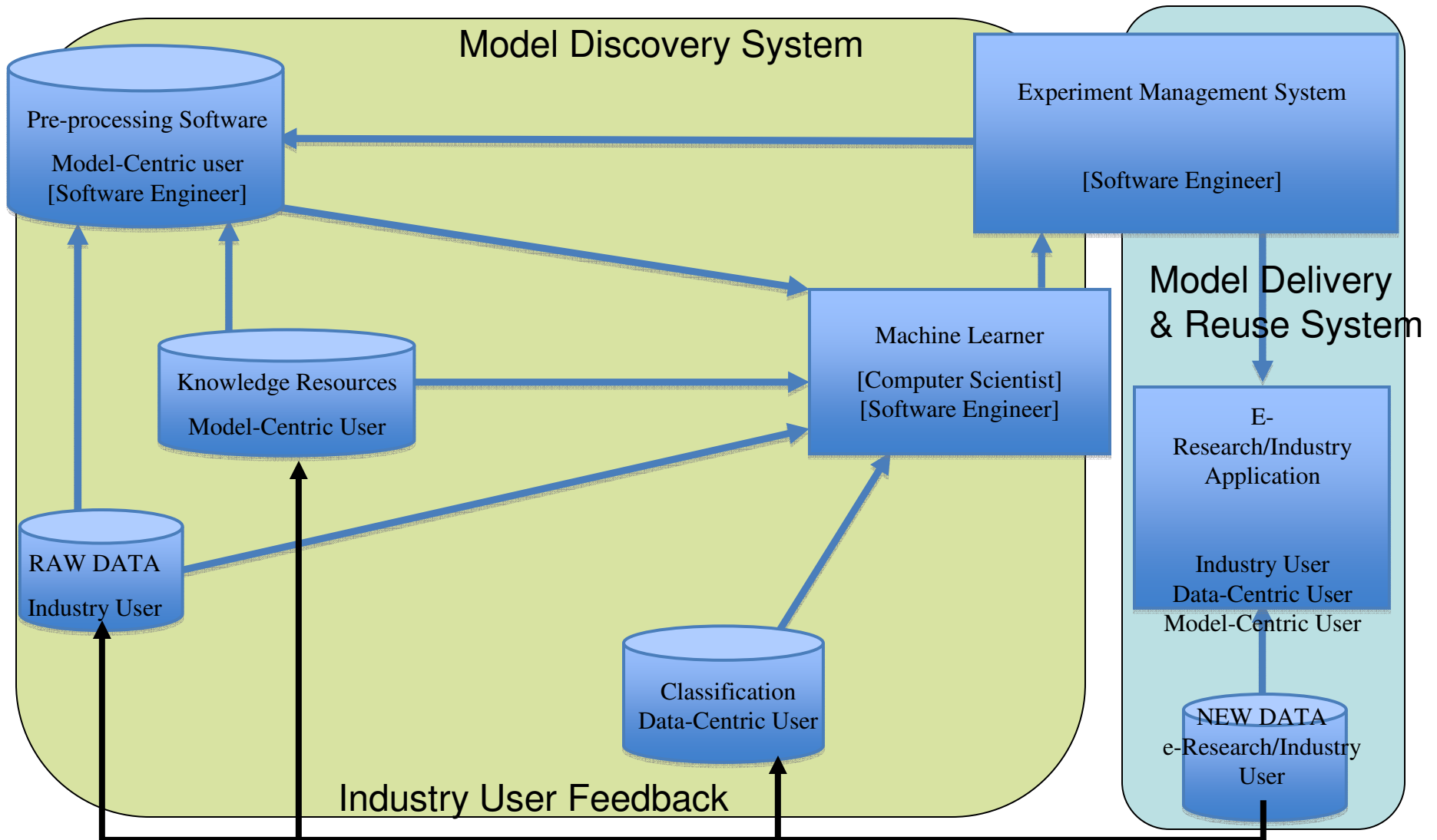


Knowledge Discovery and Knowledge Reuse

- We have a methodology for discovering unrecognisable tokens and Reusing them
- Based on human interpretation
- Designed for
 - immediate feedback
 - Multiple levels of feedback

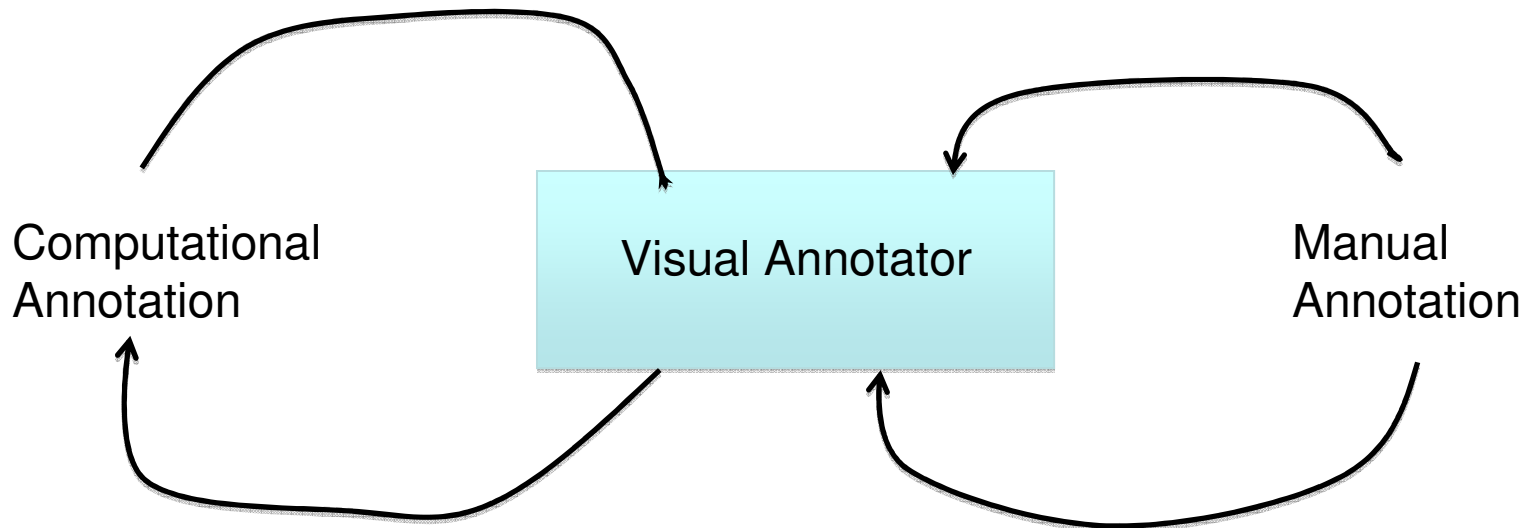


Knowledge Discovery Knowledge Reuse Methodology (KDKR)





Knowledge Discovery and Knowledge Reuse - KDKR





Knowledge Bases Used

- SNOMED CT
- Moby Lexicon
- UMLS
- Local Gazetteers of abbreviations and acronyms
- Named Entity Gazetteer
- 78889 gold standard spelling corrections



ED Corpus -Testing

- 20 million tokens
- 57,523 words
- 24,735 unknown tokens
- Tolentino et al, 2007, built a spelling corrector for mammograms using UMLS and WordNet
- This is inadequate for ED and ICU



Suggestion Generation

- 6 Rules for generation are used in a predefined sequence
- When a rule generates at least one suggestion that stops progressing through the rules.
- The following table indicates the performance of each rule used in isolation



The Rules in Usage Order

- Edit Distance 1
- Concatenated words - missing space
- Edit distance 2
- Phoentic change
- Phoenetic change Edit Distance 1
- Edit Distance 1 and concatenated words



Coverage by Rule Type

Rule	Percentage
Edit distance 1	69.8%
Two words	6.7%
Edit distance 2	85.5%
Phonetic	42.5%
Two word edit distance 1	7.1%
Multi word	6.9%
Phonetic edit distance 1	84.8%



Context Sensitive Ranking

- One word to left and right
- Cambridge Language Modelling Toolkit forms unigram, bigram, trigram distributions in the ICU corpus
- including a distribution model of the misspelt words
- Helped significantly the ranking process to get the best suggestion placed first.



Ranking Algorithm

- Word frequency models
 - Corpus frequency
 - Knowledge based frequency
 - Combination of the two
- Models for concatenated words
 - Ranking the first two words
 - Ranking first and last words



Heuristics to Improve First Suggestion Ranking

- For Edit Distance 1 rank by knowledge base frequencies, i.e. suggestions that occur more frequently in the knowledge base are more likely to occur in the corpus - reduced accuracy
- If a misspelling ended in “-ly” promote the highest “-ly” ending word to the top of the list.



Heuristics to Improve First Suggestion Ranking

- Orthographic variation was ignored - all words were converted to lower case
- Combined knowledge base and corpus frequencies removed many ties



The Data Set

- Concord Hospital, Sydney - ED
- 57,523 unique words - types
- 7442 misspellings with known corrections
- Test Data
- 12,438 words Concord ED - manually corrected but not in our KB
- 65,429 corrected misspellings from 164,302 unique words - RPAH - ICU



Experiment Results - Word List Heuristics

Rule	Training data	Concord-ED	RPAH-ICU
(Baseline)	60.62	63.78	62.28
+ 'Two words'	62.31	74.99	66.42
+ 'Phonetic'	62.31	75.11	66.57
+ 'Two word edits 1'	62.31	75.11	66.58
'Multi word'	62.31	75.11	66.61



Results - Computing Suggestions

- Baseline - KB and Edit Distance 1 & 2, 1st suggestions matched - 60.52%
- Each heuristic for computing the suggestions was added leading to a maximum of 75.11%

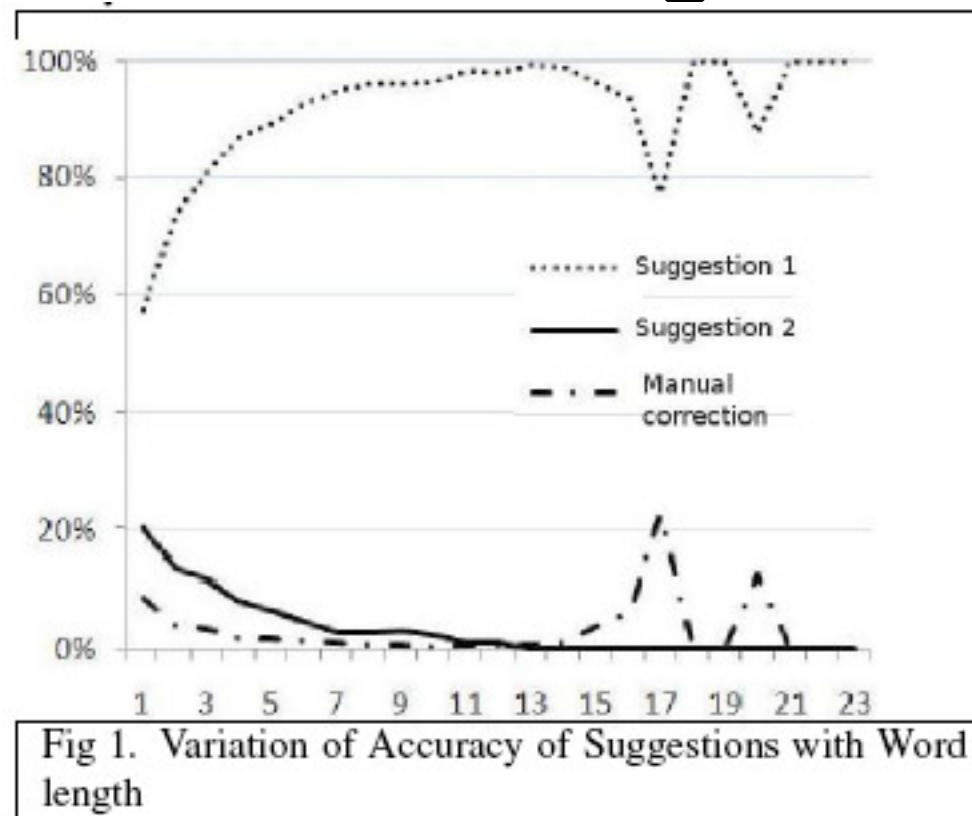


Ranking Rule Accuracies

Ranking Rule	Training Data	Concord-ED	RPAH-ICU
Corpus word frequency	87.19	93.17	81.24
Knowledge base word frequency	83.41	93.07	79.99
combined knowledge base and corpus	87.25	93.54	81.45
Phonetic + Phonetic edit distance 1	87.24	93.43	81.69
First two words ranking	87.25	92.89	81.75
First and last word	87.26	93.05	81.83



Variation of Suggestions by Word Length





Analysis

- Variation between corpora
- First-last word better than 1st-2nd word model as middle words are short and less representative
- Accuracy increased with word length
- Corpus frequencies and context probabilities helpful
- Lexica can oversupply alternatives



Accuracy Removing Lexica

Rule	Correct Num (out of 65429)	RPAH-ICU corpus
Remove UMLS	54038	82.59
Remove SCT	53769	82.18
Remove Both	54857	83.84
Remove Gazetteers	53614	81.94
Remove All	55197	84.36
Remove All + Language Model	54985	84.04



Applications

- Lexicon Management system
- Intelligent Clinical Notes System
- Visual Annotator
- END