Detecting Abbreviations Using Machine Learning Methods

Yonghui Wu[1] S. Trent Rosenbloom[1] Joshua C. Denny[1]
Randolph A. Miller[1] Subramani Mani[1] Dario A. Giuse[1]
Hua Xu[1]

Contact: yonghui.wu@Vanderbilt.Edu

[1]Department of Biomedical Informatics, Vanderbilt University

Oct 25, 2011

Challenges of Handling Clinical Abbreviations

- Pervasive use
- Highly dynamic
- Ambiguous

Example

Mr. XXX is a gentleman with a PMH sig for recent CEA , CAD , HTN, HLD , and hypothyroidism who was transferred from . . . etoh . . .

[Acronyms] , [Shorted words or phrases] , [Symbols]

Challenges of Handling Clinical Abbreviations

- Pervasive use
- Highly dynamic
- Ambiguous

Clinical abbreviations varied greatly by different { institute , type of notes and people }

Challenges of Handling Clinical Abbreviations

- Pervasive use
- Highly dynamic
- Ambiguous

Example

- Ambiguous senses: “pt – patient” and “pt – physical therapy”
- Ambiguous between abbreviations and English word: “mom – mother” and “mom–milk of magnesia”

Objective

- To determine whether a token in clinical text is an abbreviation or not.

Eg.

… gentleman with a PMH sig for recent CEA

Objective

- To determine whether a token in clinical text is an abbreviation or not.

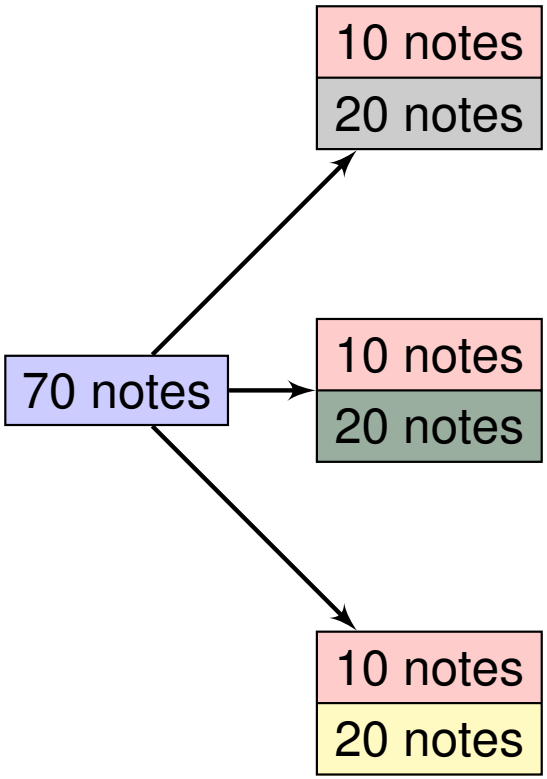
Eg.

…	gentleman	with	a	PMH	sig	for	recent	CEA
	×	×	×	√	×	×	×	√

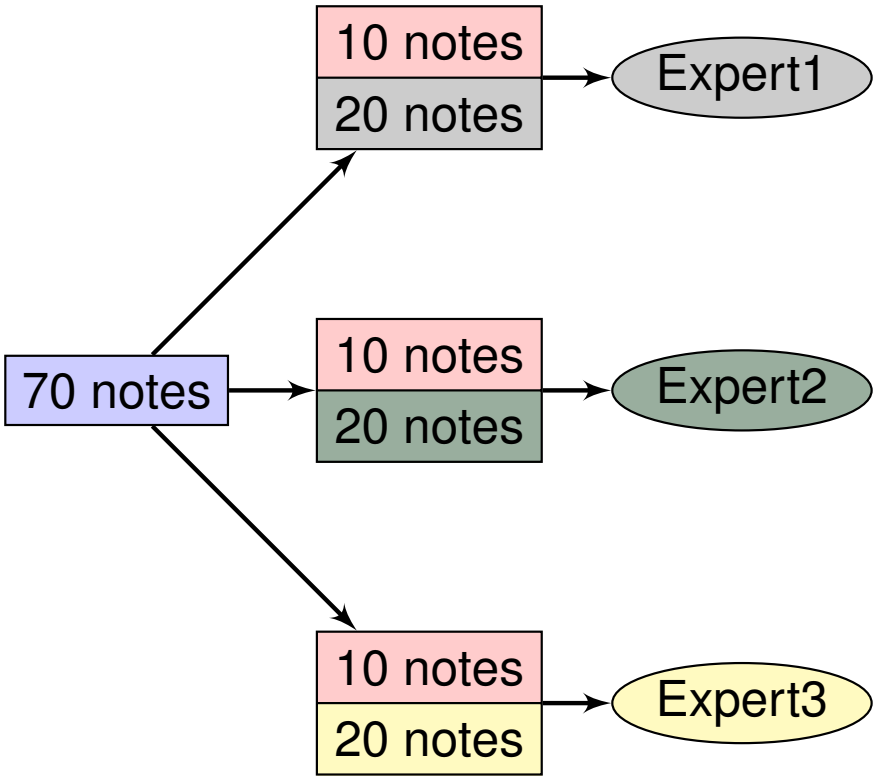
Study Design

70 notes

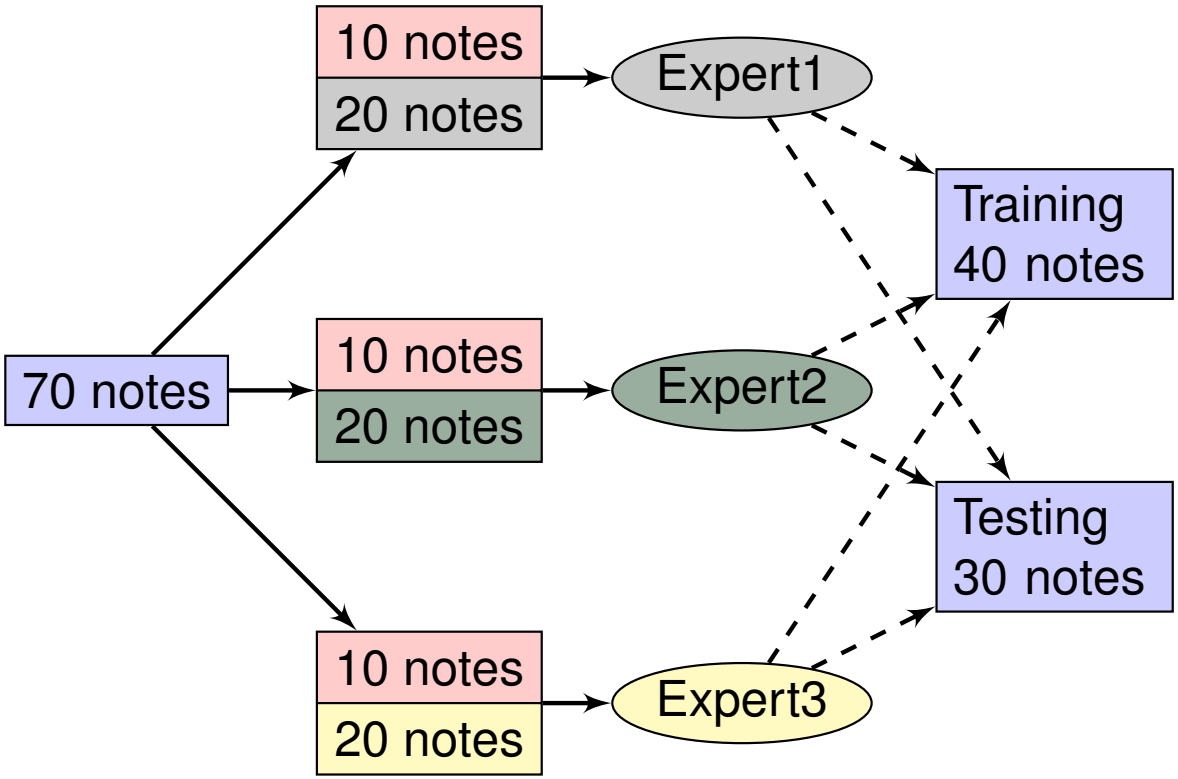
Study Design



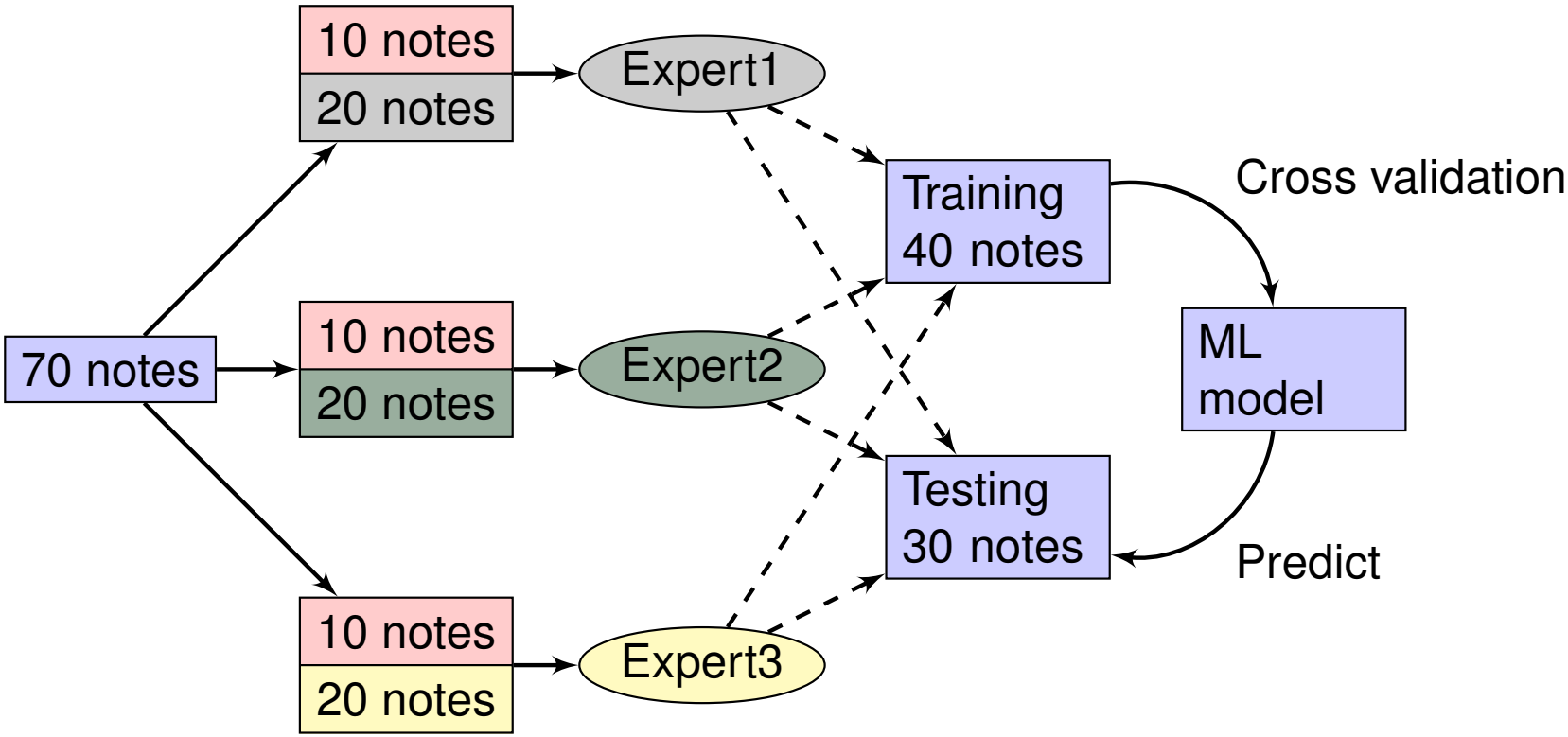
Study Design



Study Design



Study Design



Baseline & ML methods

- Baseline: dictionary based method
- ML methods:
  - Decision Tree (DT)
  - Random Forests (RF)
  - Support Vector Machines (SVMs)

Baseline & ML methods

- Baseline: dictionary based method
- ML methods:
  - Decision Tree (DT)
  - Random Forests (RF)
  - Support Vector Machines (SVMs)

- Combined Methods
  - Scheme 1: At least one of the ML methods predict “yes”
  - Scheme 2: At least two of the ML methods predict “yes”
  - Scheme 3: All ML methods predict “yes”

Features

- Groups of features
  1. **Word formation features**: Uppercase, dash, slash, dot, alphabetic/numeric characters, length of the word, misspelling features
  2. **Vowel/consonant features**
  3. **Knowledge bases feature**: English/medical word
  4. **Corpus features** (Average Word Frequency)
  5. **Context features** (The features derived from the token before and after the current word)
- Feature selection: select the best combination of features using the 10-fold cross validation on training set.

Evaluation

- Precision, Recall and F-score
- Three levels of evaluation: ALL, UNIQUE and UNKNOWN
- Two gold-standard: GS-1 and GS-2

Precisin, Recall and F-score

- $Precision = \frac{True\ System\ Predicted\ Abbreviations}{All\ System\ Predicted\ Abbreviations}$
- $Recall = \frac{True\ System\ Predicted\ Abbreviations}{All\ Experts\ Labeled\ Abbreviations}$
- $F\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall}$

Evaluation

- Precision, Recall and F-score
- Three levels of evaluation: ALL, UNIQUE and UNKNOWN
- Two gold-standard: GS-1 and GS-2

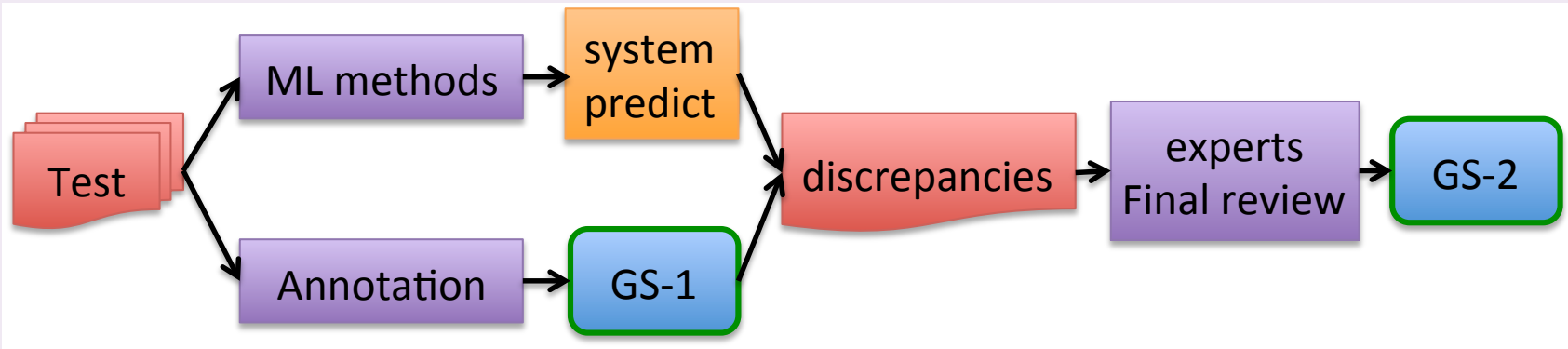
Three levels of evaluation

- *ALL*: Count duplicate abbreviations separately.
- *UNIQUE*: Considered each individual abbreviation only.
- *UNKNOWN*: Remove the abbreviations occurred in training set.

Evaluation

- Precision, Recall and F-score
- Three levels of evaluation: ALL, UNIQUE and UNKNOWN
- Two gold-standard: GS-1 and GS-2

Two Gold standard



Description of data set

Inter–Annotator Agreement

- $KAPPA = 0.886$

Description of data

	Number of Documents	Positive		Negative	
		All	Unique	All	Unique
Training	40	1,386	475	16,839	3,736
Test	30	1,402	448	12,511	3,060

Cross Validation Using Different feature sets

Description of feature sets

1. word formation features;
2. vowel feature;
3. knowledge-based features;
4. word frequency (global feature);
5. context features derived from the previous/next token.

Methods		1	1+2	1+2+3	1+2+3+4	1+2+3+4+5
DT	Precision	78.5%	78.2%	90.8%	90.4%	90.1%
	Recall	53.6%	71.0%	80.1%	91.7%	89.6%
	F-score	63.7%	74.4%	85.1%	91.0%	89.9%
RF	Precision	79.4%	78.5%	92.3%	93.1%	93.2%
	Recall	54.1%	73.0%	83.2%	94.4%	92.3%
	F-score	64.4%	75.7%	87.5%	93.7%	92.8%
SVM	Precision	67.7%	70.1%	83.2%	89.1%	93.5%
	Recall	62.0%	80.8%	90.7%	92.7%	74.7%
	F-score	64.7%	75.1%	86.8%	91.0%	83.1%

Cross Validation Using Different feature sets

Description of feature sets

1. word formation features;
2. vowel feature;
3. knowledge-based features;
4. word frequency (global feature);
5. context features derived from the previous/next token.

Methods		1	1+2	1+2+3	1+2+3+4	1+2+3+4+5
DT	Precision	78.5%	78.2%	90.8%	90.4%	90.1%
	Recall	53.6%	71.0%	80.1%	91.7%	89.6%
	F-score	63.7%	74.4%	85.1%	91.0%	89.9%
RF	Precision	79.4%	78.5%	92.3%	93.1%	93.2%
	Recall	54.1%	73.0%	83.2%	94.4%	92.3%
	F-score	64.4%	75.7%	87.5%	93.7%	92.8%
SVM	Precision	67.7%	70.1%	83.2%	89.1%	93.5%
	Recall	62.0%	80.8%	90.7%	92.7%	74.7%
	F-score	64.7%	75.1%	86.8%	91.0%	83.1%

Predict Result on Test Set (GS-2)

Method		Precision	Recall	F-score
DT	All	98.0%	89.8%	93.7%
	Unique	95.2%	82.9%	88.6%
	Unknown	90.8%	78.3%	**84.1%**
RF	All	98.8%	91.2%	**94.8%**
	Unique	96.7%	81.8%	88.7%
	Unknown	93.1%	71.2%	80.7%
SVM	All	98.1%	91.3%	94.5%
	Unique	97.5%	82.9%	**89.6%**
	Unknown	94.8%	74.1%	83.2%
Baseline	All	86.8%	86.0%	86.4%
	Unique	73.7%	87.3%	79.9%

Predict Result on Test Set (GS-2)

Method		Precision	Recall	F-score
DT	All	98.0%	89.8%	93.7%
	Unique	95.2%	82.9%	88.6%
	Unknown	90.8%	78.3%	**84.1%**
RF	All	98.8%	91.2%	**94.8%**
	Unique	96.7%	81.8%	88.7%
	Unknown	93.1%	71.2%	80.7%
SVM	All	98.1%	91.3%	94.5%
	Unique	97.5%	82.9%	**89.6%**
	Unknown	94.8%	74.1%	83.2%
Baseline	All	86.8%	86.0%	86.4%
	Unique	73.7%	87.3%	79.9%

Voting between SVM, RF, and DT

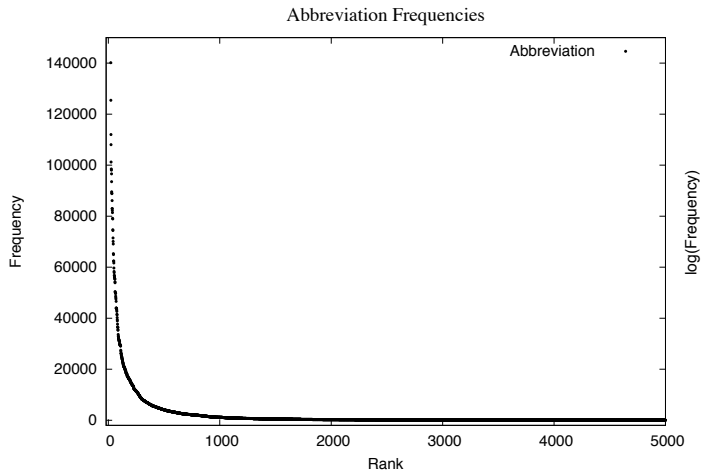
Schemes		Precision	Recall	F-score
Scheme 1	All	97.0%	94.5%	**95.7%**
	Unique	93.0%	90.0%	**91.6%**
	Unknown	86.9%	85.2%	**86.1%**
Scheme 2	All	98.6%	92.0%	95.2%
	Unique	98.2%	83.2%	90.1%
	Unknown	94.4%	75.8%	84.1%
Scheme 3	All	99.5%	85.8%	92.2%
	Unique	98.5%	73.8%	84.4%
	Unknown	96.7%	61.5%	75.5%

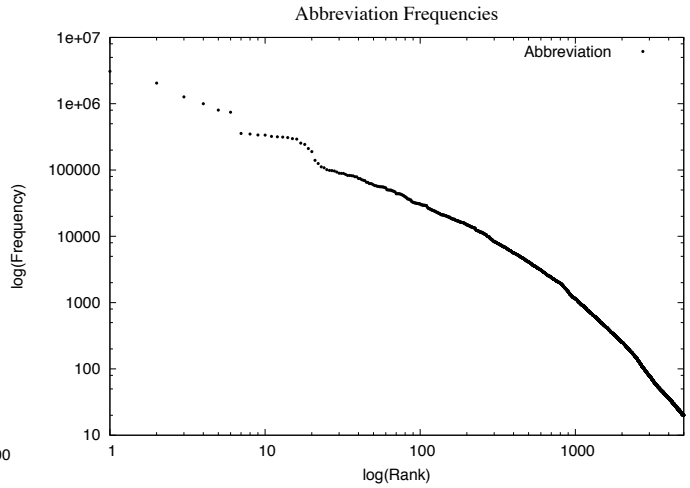
Voting between SVM, RF, and DT

Schemes		Precision	Recall	F-score
Scheme 1	All	97.0%	94.5%	**95.7%**
	Unique	93.0%	90.0%	**91.6%**
	Unknown	86.9%	85.2%	**86.1%**
Scheme 2	All	98.6%	92.0%	95.2%
	Unique	98.2%	83.2%	90.1%
	Unknown	94.4%	75.8%	84.1%
Scheme 3	All	99.5%	85.8%	92.2%
	Unique	98.5%	73.8%	84.4%
	Unknown	96.7%	61.5%	75.5%

Apply to large corpus

- Data: 10 years' discharge summaries: 560,650
- Predicted result: 26,257 abbreviations with 24,120,704 occurrences
- Top abbreviations in discharge notes:
  mg (frequency : 2,049,019), po (1,265,985), md (1,001,341), dr (801,919), prn (356,486), mrn (349,093), bid (321,122), iv (313,744), pt (308,048)





Discussion

Discussion

- For domain experts, annotating abbreviations is not hard ,but not error-free task. ($KAPPA = 0.8864$).

Discussion

- For domain experts, annotating abbreviations is not hard ,but not error-free task. ($KAPPA = 0.8864$).
- Ambiguity between an abbreviation and an English word is a bottle neck.

Discussion

- For domain experts, annotating abbreviations is not hard ,but not error-free task. ($KAPPA = 0.8864$).
- Ambiguity between an abbreviation and an English word is a bottle neck.
- Detecting unknown abbreviations is more challenging (Best F-score: 86.1%).

Discussion

- For domain experts, annotating abbreviations is not hard ,but not error-free task. ($KAPPA = 0.8864$).
- Ambiguity between an abbreviation and an English word is a bottle neck.
- Detecting unknown abbreviations is more challenging (Best F-score: 86.1%).
- Limitation: handle multi-word abbreviations.

Conclusion & Future Work

Conclusion

- The ML methods coulde be applied to clinical corpora to create a useful lexical source of abbreviations.
- Voting between ML methods improved system performance (Scheme-1: f-score=$\%95.7$)

Conclusion & Future Work

Conclusion

- The ML methods coulde be applied to clinical corpora to create a useful lexical source of abbreviations.
- Voting between ML methods improved system performance (Scheme-1: f-score=$\%95.7$)

Future Work

- Create a lexical source of abbreviations.
- Build corpus-based sense inventory.
- Develop a on-line abbreviations handling system that could automatically detect and disambiguate abbreviations in real time.

Acknowledgement

- Grant: NLM R01LM010681 (**PI – Xu**)
- Lexical sources of abbreviations from: UMLS LRABR, ADAM and Berman's abbreviation list.
- Help from: *S. Trent Rosenbloom,Joshua C. Denny, Randolph A. Miller, Subramani Mani, Dario A. Giuse, Hua Xu*

Thanks

Q & A

Email: yonghui.wu@Vanderbilt.Edu