MINING USER INTENTIONS FROM MEDICAL QUERIES: A NEURAL NETWORK BASED HETEROGENEOUS JOINTLY MODELING APPROACH

Source: WWW'16

Advisor: Jia-Lin,Koh

Speaker: Ming-Chieh,Chiang

Date: 2017/12/05

Outline

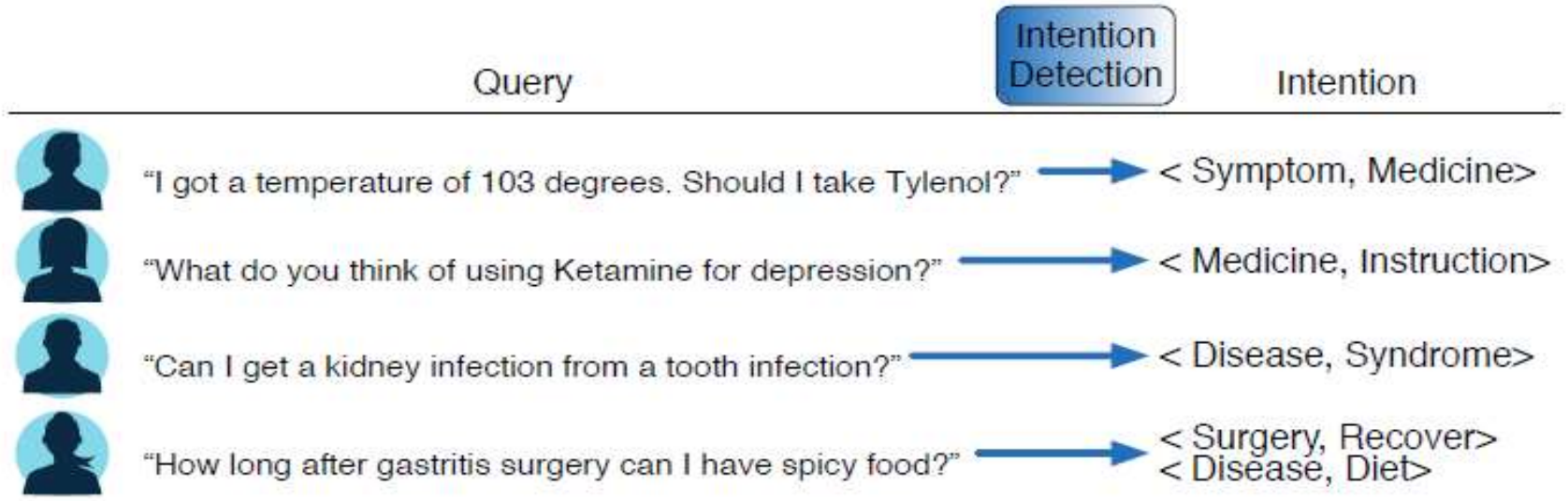
- **Introduction**
- Method
- Experiment
- Conclusion

Introduction

- Motivation
  - Text queries are naturally encoded with user intentions
  - Words from different topic categories tend to co-occur in medical related queries
  - This work aims to discover user intentions from medical-related text queries that users provided online

Introduction

- Goal
  - Input : medical query
  - Output : intentions

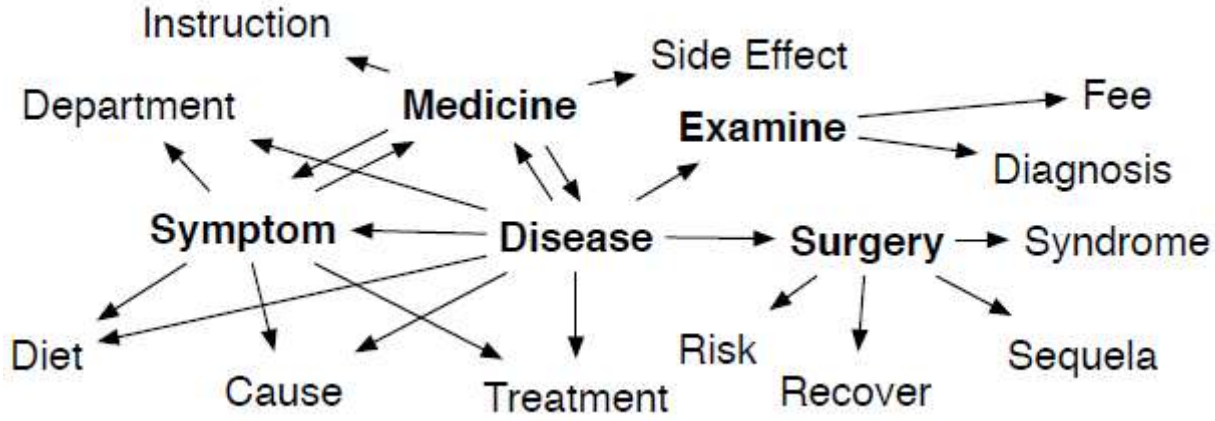


Introduction

- Definition of intention

$$I = \{< s, n >\}$$

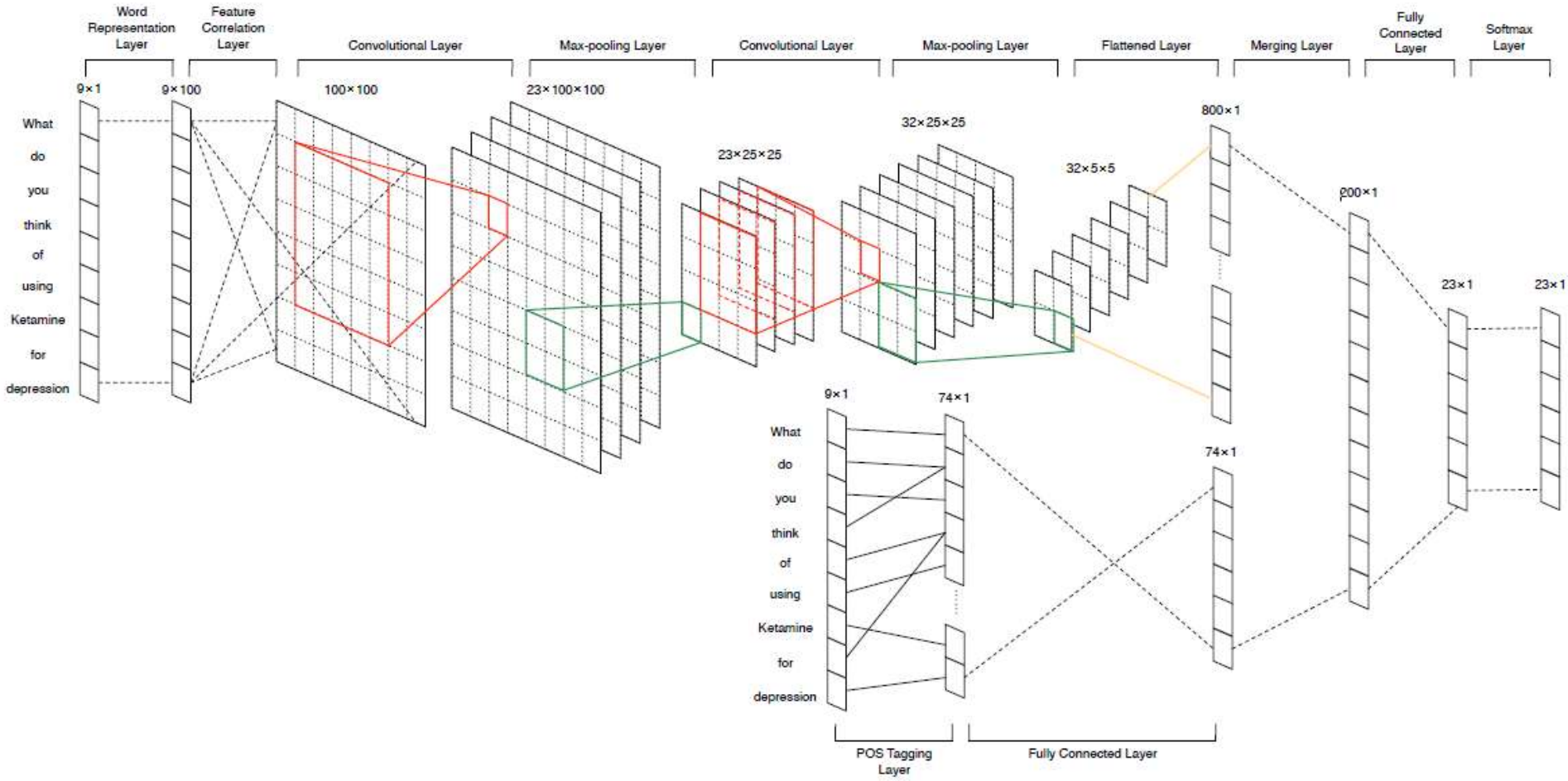
  - By describing related information in concept s, the user is looking for corresponding information about concept n.



Outline

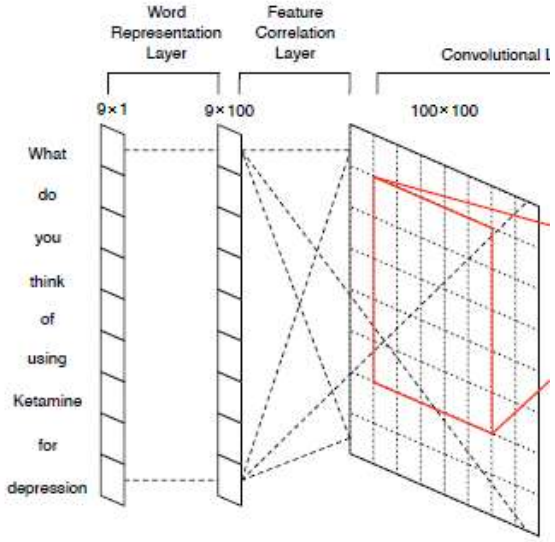
- Introduction
- **Method**
- Experiment
- Conclusion

Architecture



Feature-level modeling

- Pairwise feature correlation matrix



  - sim(Mi,Mj) : the similarity between feature Mi and Mj

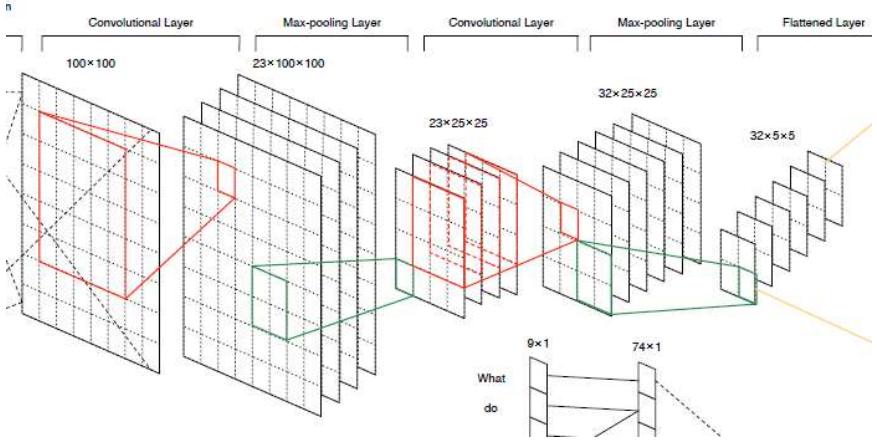
$$S = \begin{bmatrix} sim(M_1, M_1) & sim(M_1, M_2) & \cdots & sim(M_1, M_m) \\ sim(M_2, M_1) & sim(M_2, M_2) & & sim(M_2, M_m) \\ \vdots & \vdots & & \vdots \\ sim(M_m, M_1) & sim(M_m, M_2) & \cdots & sim(M_m, M_m) \end{bmatrix}$$

Feature-level modeling

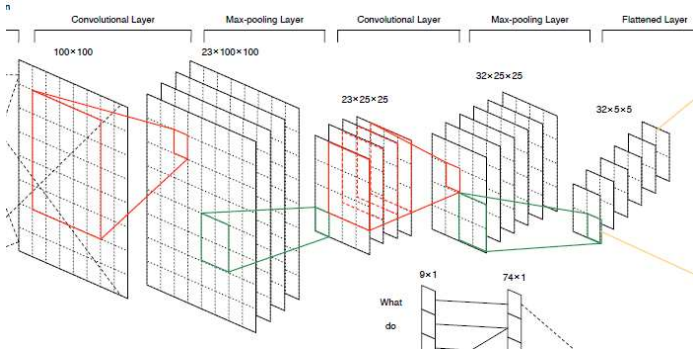
- Convolution operation

$$c = f(\mathbf{t}_k \cdot \mathbf{x} + b_k)$$

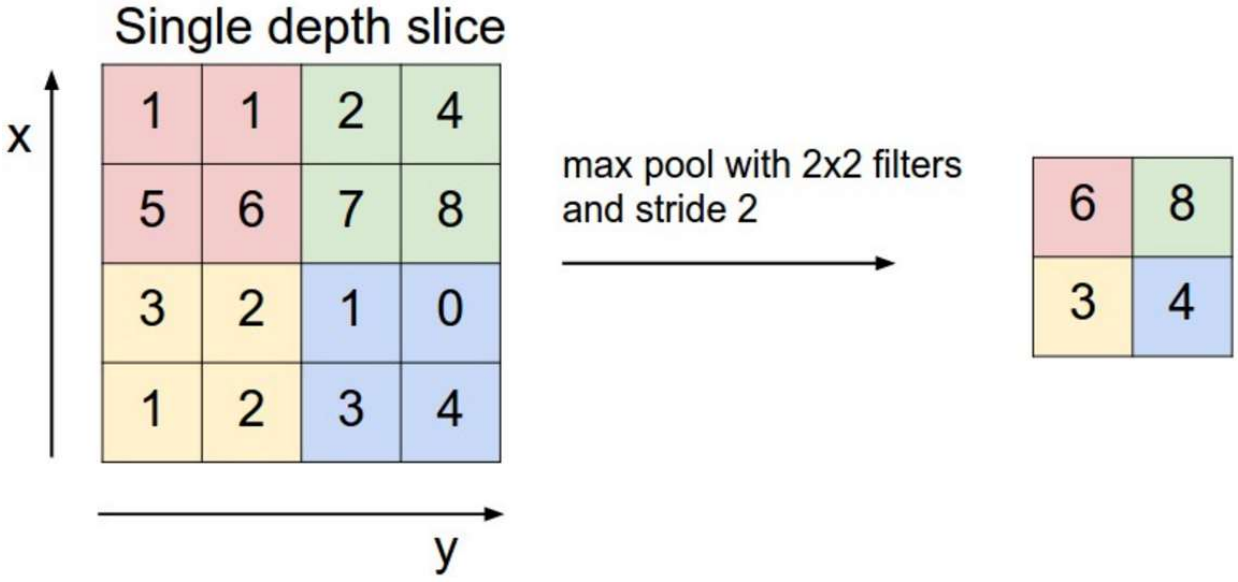
  - k filters
  - tk : weight matrix
  - x : convolution region
  - bk : bias
  - f : ReLU(x) = max(0,x)



Feature-level modeling

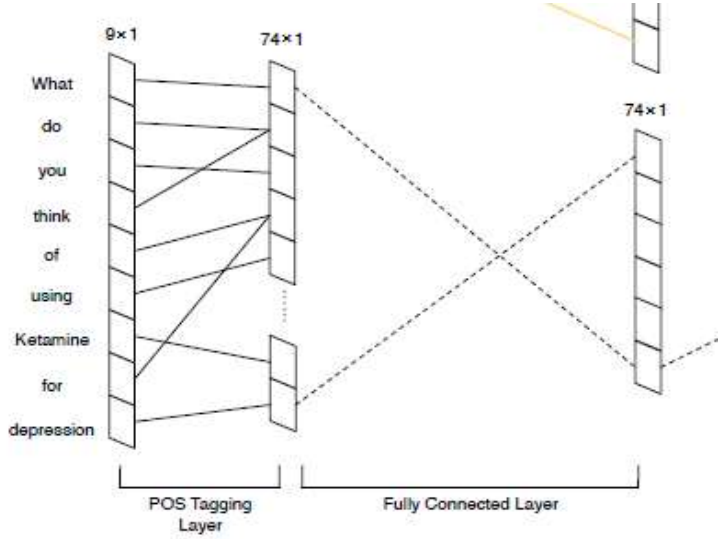


- Pooling operation
  - a subsampling function that returns the maximum of a set of values



POS tagging

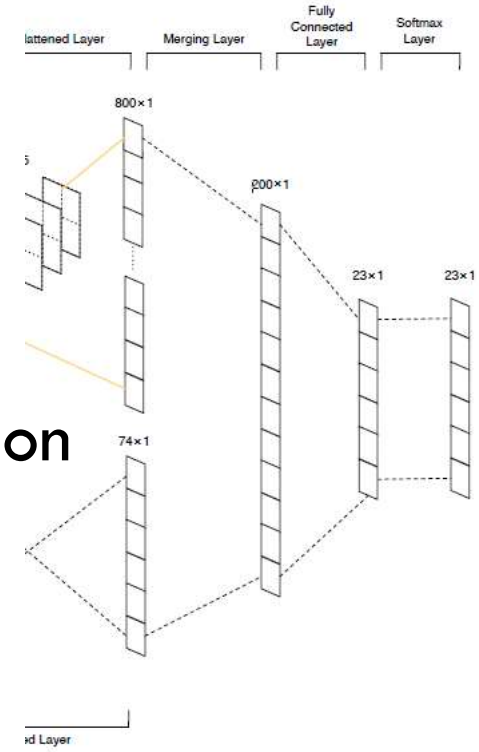
- POS tagging is used as word categories
- Calculate the number of occurrence of each tag
- Fully connected layer : estimate the contribution of different POS tags



Jointly modeling

- To overcome the domain coverage challenge.
- “ I have been taking Tylenol .”
- “ I have been taking aspirin”
- Tylenol & aspirin :

  the word category is “n-medicine”
- Concatenate results and reduce dimension



Increasing model generalization ability

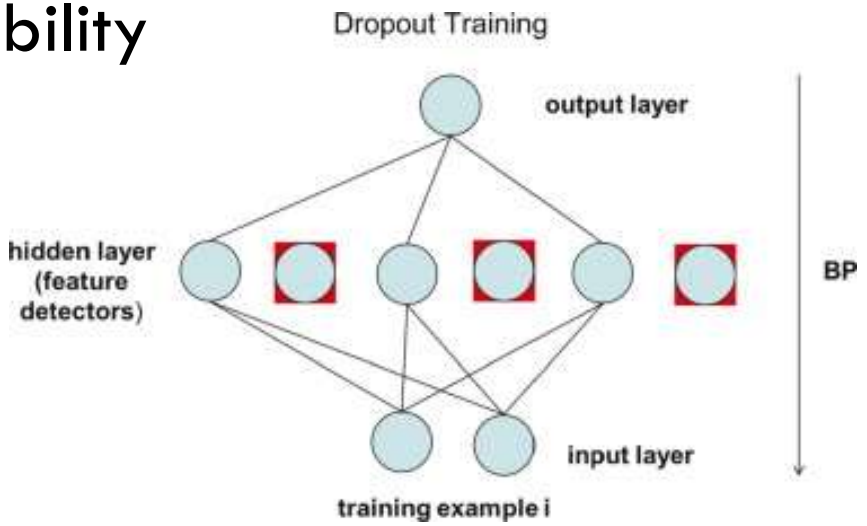
- Data augmentation
  - To reduce overfitting
  - Sentence Rephrasing
  - Use the nearest neighbors of a word in a vector space to generate candidate rephrasing words
  - Constrain original word and candidate words with a equality constraint on POS type as well as similarity constraints

Increasing model generalization ability

- Data augmentation
  - Calculate the nearest neighbors of words
  - Check each candidate word that whether it has the same tag with each word
  - Use threshold for the similarity measurement
  - If the new word meets those constrains, then replacing this old word by the candidate word to generate a new query

Increasing model generalization ability

- Dropout
  - A regulation method to overcome co-adapting of feature detectors
  - To reduce test error
  - Dropout layer is applied after each pooling layer with 0.5 probability



Outline

- Introduction
- Method
- **Experiment**
- Conclusion

Dataset

- corpus : http://club.xywy.com/
- 64 million records
- Pre-processing : word segmentation
- Use word2vec to train vector representation of words
- The vectors have dimensionality of 100 and were trained using the Skip-gram
- Window size : 8
- Minimum occurrence count : 5

Baseline methods

- SVM-FC (Feature-level Correlation)
- LR-FC (Logistic Regression)
- NNID-ZP (Zero Padding)
- NNID-FC
- NNID-JM (Jointly Modeling)
- NNID-JMSR (Sentence Rephrasing)

Performance

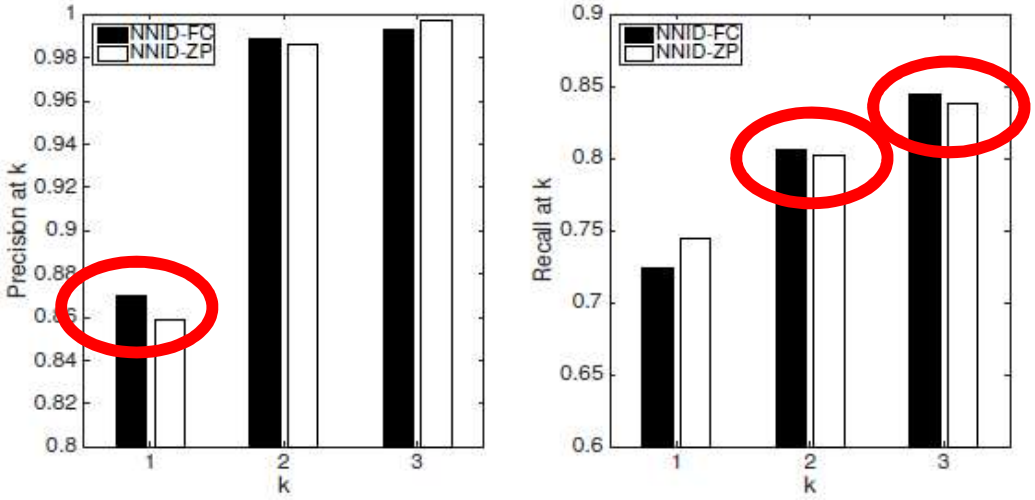


Figure 6: NNID-FC vs NNID-ZP.

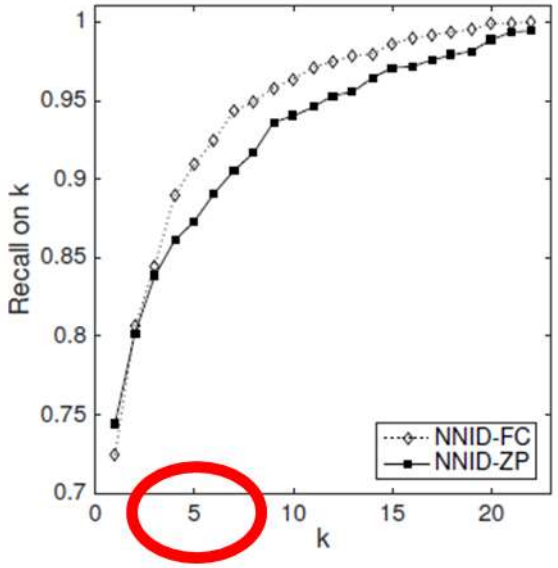


Figure 7: Recall of NNID-FC vs NNID-ZP.

Performance

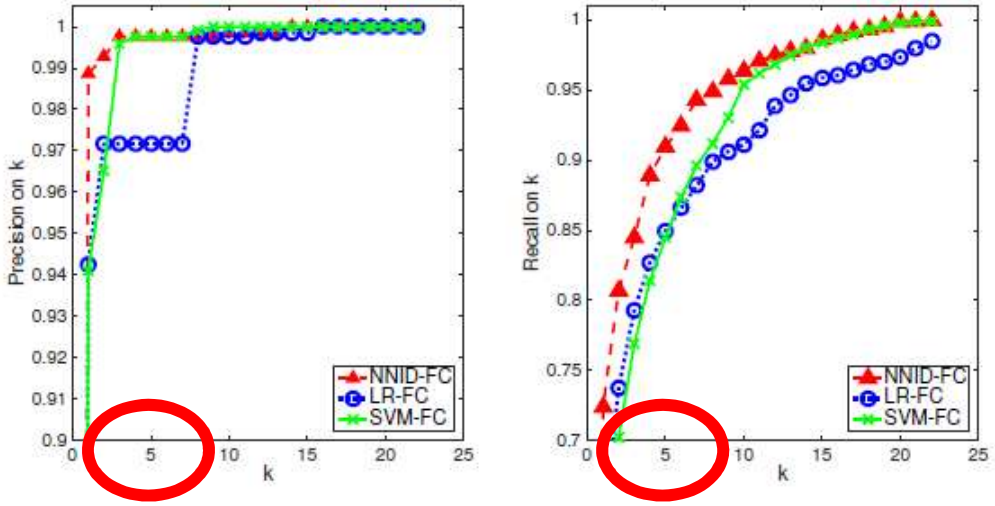


Figure 8: NNID-FC vs LR-FC vs SVM-FC

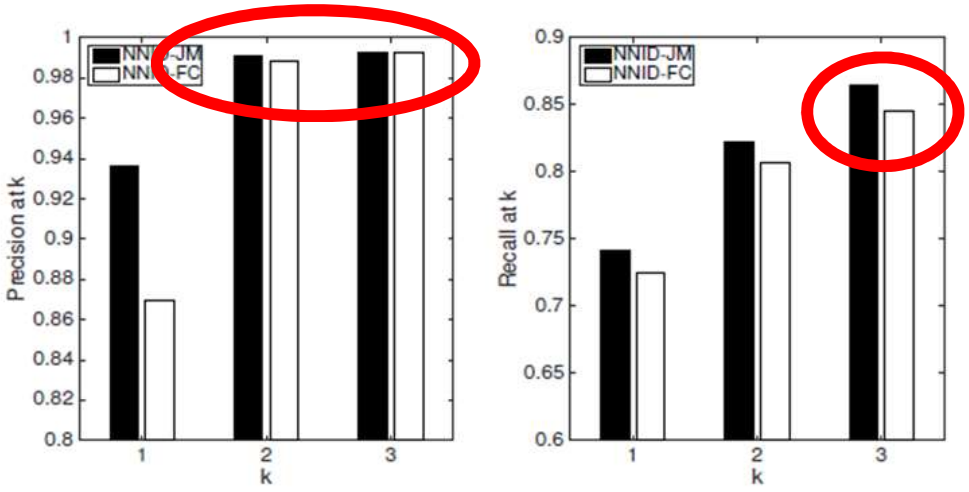


Figure 9: NNID-JM vs NNID-FC.

Performance

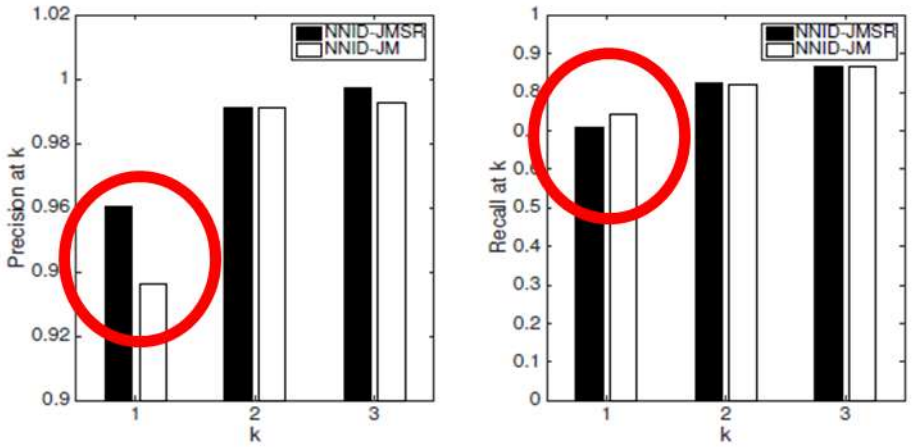


Figure 10: NNID-JMSR vs NNID-JM.

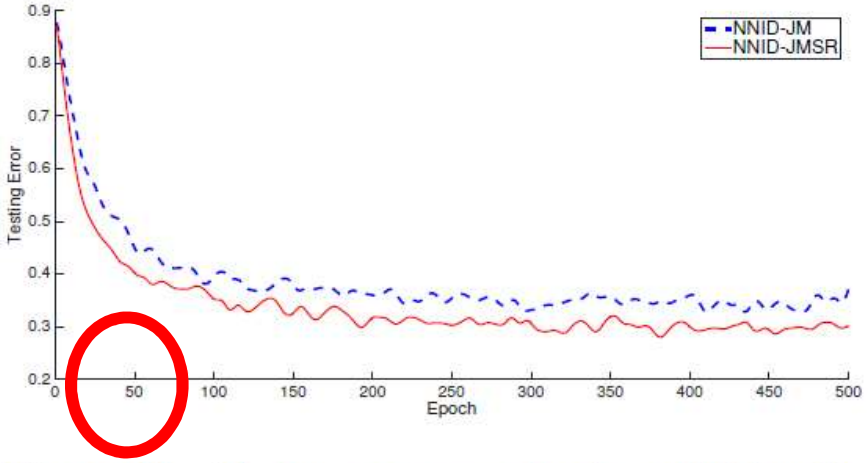


Figure 11: Testing error on top-1 prediction for NNID-JMSR and NNID-JM.

Case

•**Query**: How much does it costs for a Lumbar CT? Recently my lumbar always hurts. (腰椎CT检查大概需要多少费用? 最近后腰老是酸疼。)

Prediction:

Rank	Intention	Probability
1	<examine,fee>	0.986955
2	<symptom,examine>	0.012433
3	<symptom,department>	0.000475
4	<disease,department>	8.50e-05
5	<examine,diagnose>	3.51e-05

Outline

- Introduction
- Method
- Experiment
- **Conclusion**

Conclusion

- Intention detection for medical query will provide a new opportunity to connect patients with medical resources more seamlessly both in physical world and on the WWW
- Present a jointly modeling approach to model intentions that users encoded in medical related text queries
- The method can be generalized and integrated into other existing applications as well