Mapping Chinese Medical Records to UMLS

Fan Meng, Ph.D.
Psychiatry Department and
Molecular and Behavioral Neuroscience Institute
University of Michigan

University of Michigan Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

Motivating Use Case

- Compare US and Chinese emergency medicine department patients survival rate: using patient symptom descriptions to predict patient survival time through machine learning.
- Requires the appropriate mapping of different symptom description terms representing the same concept in both languages.
- Translation provided by third party companies at a cost of $1/term did not provide satisfactory results

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

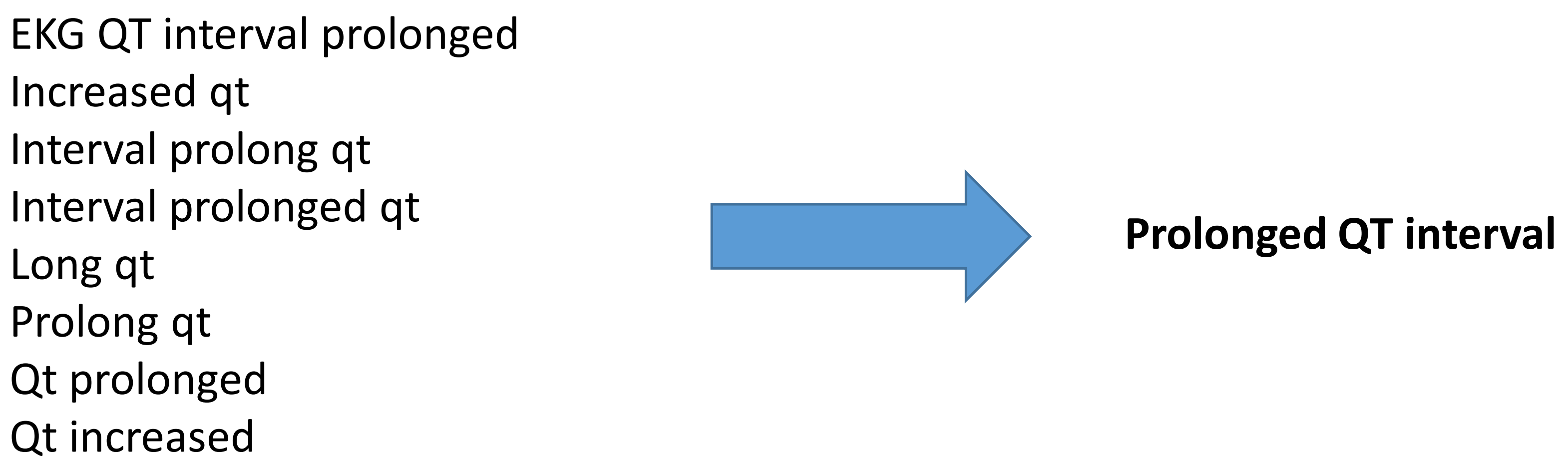
北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

Unified Medical Language System (UMLS)

- UMLS is a is a compendium of many controlled vocabularies in the biomedical sciences. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems.
- The US National Library of Medicine developed UMLS in 1986 and has been updating UMLS at twice a year frequency recently.
- Key Statistics for 2019AA released:
  - Unique concepts: 3,848,696
  - Concept names: 14,608,809
  - Data sources : 210

An Example of One Concept Associated with Multiple Terms

The same disease, medication, procedure, etc. can be described by different words, phrases and in medical records and literature.



EKG QT interval prolonged
Increased qt
Interval prolong qt
Interval prolonged qt
Long qt
Prolong qt
Qt prolonged
Qt increased

Prolonged QT interval

Effective medical record/literature mining requires the linking of different expressions to the same concepts first.

UMLS Concept: Prolonged QT interval

EKG QT interval prolonged
Increased qt
Interval prolong qt
Interval prolonged qt
Long qt
Prolong qt
Qt prolonged
Qt increased
…
心電図ＱＴ延長
ＱＴ延長
ＱＴ間隔延長
QTｴﾝﾁｮｳ
QTｶﾝｶｸｴﾝﾁｮｳ
…

→

Prolonged QT interval
(C0151878)

University of Michigan Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

Challenges for Mining of Chinese Medical Records Using UMLS

1. Chinese medical terms are only 0.54% of UMLS concept names

Language	Name Count	% of Metathesaurus	Rank
ENG	10,406,797	71.24%	1
JPN	325,365	2.23%	4
CHI	78292	0.54%	12

2. Even if there are Chinese translation (e.g., ICD10 by scholars in Taiwan), only the primary name for the concept is translated, there is no list of alternative Chinese terms that should be linked to the related concepts.

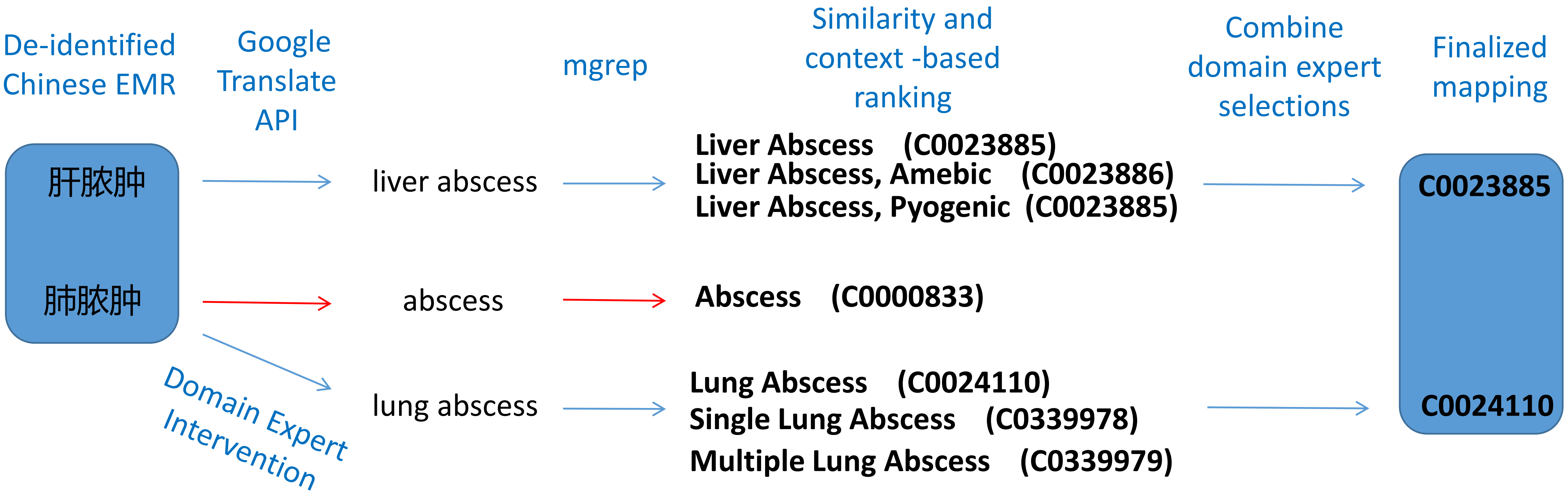
3. Simple dictionary based translation does not work well: context is important

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

Chinese EMR to UMLS Term Mapping: Overview



De-identified Chinese EMR

Google Translate API

mgrep

Similarity and context -based ranking

Combine domain expert selections

Finalized mapping

肝脓肿

肺脓肿

liver abscess

abscess

Domain Expert Intervention

lung abscess

Liver Abscess (C0023885)
Liver Abscess, Amebic (C0023886)
Liver Abscess, Pyogenic (C0023885)

Abscess (C0000833)

Lung Abscess (C0024110)
Single Lung Abscess (C0339978)
Multiple Lung Abscess (C0339979)

C0023885

C0024110

University of Michigan Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部

PEKING UNIVERSITY HEALTH SCIENCE CENTER

1912

Main Considerations

- Translating ontologies into Chinese:
    - Chinese synonyms unknown
    - Few medical ontologies are in Chinese
    - Chinese concept mapping solutions ?
- Translating Chinese medical records into English:
    - Ontologies and mapping tools are more mature
    - Ideal for cross-country comparison
- Cost of medical term translation
    - AI translators are low cost but far from satisfactory
    - Translation service companies are expensive and not good at medical terms

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

Our Strategy: AI Translation+ Cloud Sourcing

- Chinese terms are sent to 5 AI Translation solutions (Google, Bing, Baidu, Tecent, and Youdao) to obtain the initial translation results.

- The best AI translation for each term is selected by medical students and doctors at PKUHSC through a custom mobile application. Results from multiple cloud sourcing participants are merged and the a dictionary is built for best translation.

- The English translation will be mapped to the UMLS concepts by mgrep, which is an extremely fast concept mapping program developed by Manhong Dai in our group. It is the default concept mapping engine behind the National Center for Biomedical Ontology Bioportal (http://bis.bioportal.bioontology.org/).

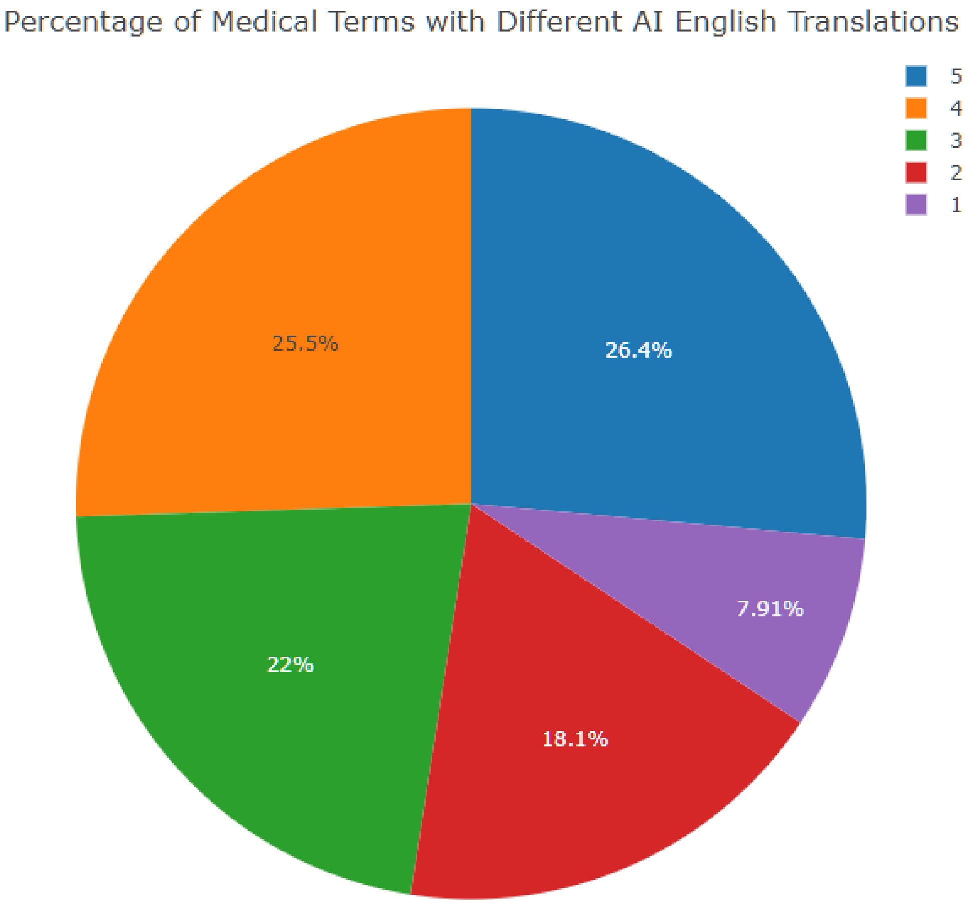
- Once the expert validated concept mapping results are obtained for a given clinical area, millions of medical records can be processed automatically.

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

AI Translation Engines Are Not Optimized for Medical Terms


Percentage of Medical Terms with Different AI English Translations

University of Michigan Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部 PEKING UNIVERSITY HEALTH SCIENCE CENTER 1912

Consistency Among AI Translators (July 2019)*

Consistency_Count	Baidu	**Google**	Microsoft	Tecent	Youdao
5	7.9%	**7.9%**	7.9%	7.9%	7.9%
4	12.9%	**13.6%**	7.1%	12.1%	11.6%
3	13.7%	**15.7%**	8.1%	11.6%	11.3%
2	18.7%	**19.3%**	12.0%	16.2%	15.3%
1	46.8%	**43.5%**	64.9%	52.3%	53.9%

* Based on 5272 medical terms

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

北京大學醫學部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

Cloud-Sourcing Web Application Design

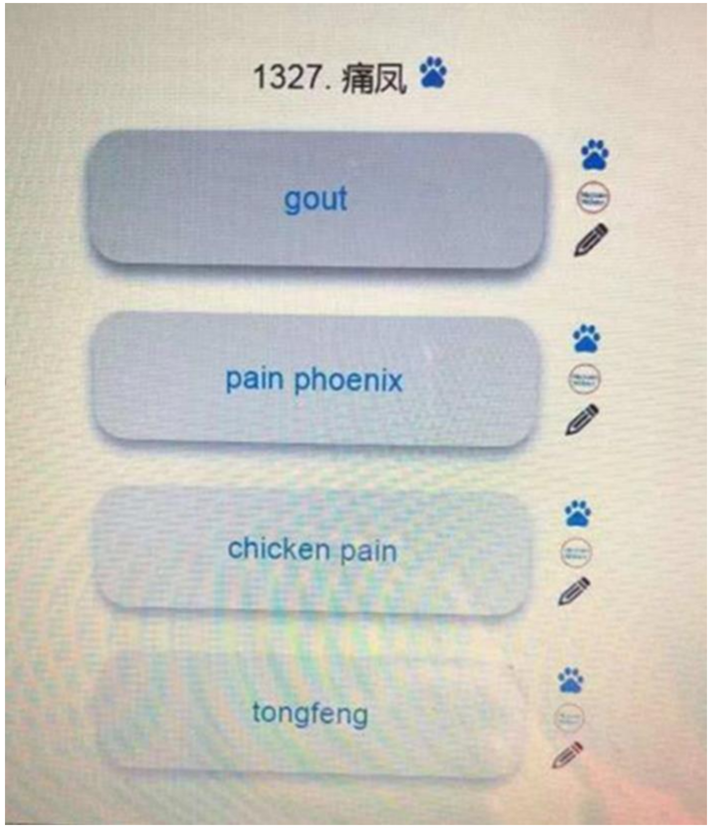
- The web application can be used across different internet connected platforms, including smartphones.
- Users can search online dictionaries for both Chinese and English terms.
- AI-translated English terms can be modified by users.
- Users is limited to go back only one term for correcting errors.
- We preset the number of answers needed for each round and typically invite 2x more participants needed to finish all terms.
- The progress of the project is displayed in real time.

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

Web Application for Cloud Sourcing-Based Validation



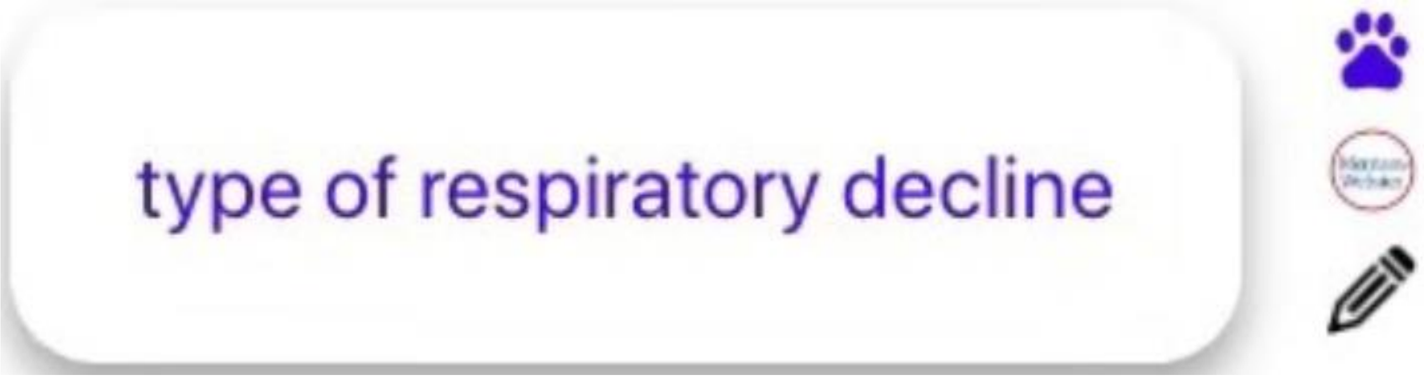
1327. 痛风

gout

pain phoenix

chicken pain

tongfeng



type of respiratory decline

University of Michigan Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部 PEKING UNIVERSITY HEALTH SCIENCE CENTER 1912

Evaluation of Cloud-Sourcing Results

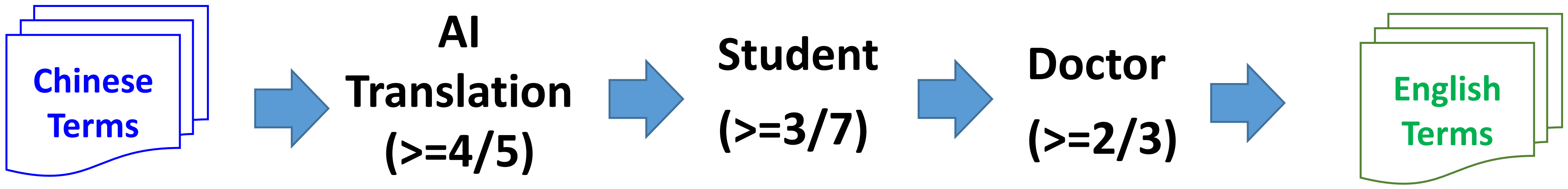
- All presented evaluations are based on answer consistency among cloud sourcing participants, not a golden standard.
- The answer selected by the majority of participants is considered to be correct.
- We also accept multiple translations for the same Chinese terms if a translation is selected by >=3 participants although not reach majority.
- Since user requested longer automatic log-out time , we do not include answers that take longer than 10 min to finish in average speed estimations.
- Our goal is to reach 90%-95% accuracy in Chinese to English translation in medical terms.

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

A 3-Tier Crowd-Sourcing Design Example



Chinese Terms

AI Translation (>=4/5)

Student (>=3/7)

Doctor (>=2/3)

English Terms

5272

1173/5272

3141/4099

622/958

4936/5272

Purely based on consensus selection, we can obtain translations for 93.6% input terms
Even if only half of the manual translations are correct for the remaining terms and consensus still has a few percent errors, we can easily achieve 90%-95% accuracy

University of Michigan Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部 PEKING UNIVERSITY HEALTH SCIENCE CENTER 1912

Cloud Sourcing Answer Type and Time Consumed

	Selection time/answer	Avg. num. of Selections/answer	Manual translation time/answer	% of Manual Translation	Total Answers	Manual Translations
R1_student	15.3	3.14	53.4	7.8%	18999	1484
R1_doctor	19.4	3.14	55	9.4%	10555	992
R2_student	26.9	3.96	62.6	9.5%	28693	2728
R2_doctor	43.7	3.73	149	4.0%	2874	115
Expert*	20	1.3	60	10%		

* Estimation, not from direct data

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

Time Estimation for Building10K Chinese Medical Term Thesaurus

Participant type (Num. of answer set)	Number of terms	Number of answers	Time/answer (second)	Total time / answer set (Hour)
AI (5)	10000	50000	1	2.78
Student (7)	8000	56000	30	93.33
Doctor(3)	2000	6000	55.4	18.47
Expert (1) *	10000	10000	24	13.33

* Estimation, not from direct data

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

Cost Estimation for Building 10K Chinese Medical Term Thesaurus

Participant type (num. of answers)	Number of terms	Total num. of answers	Cost/term (yuan)	Total cost (Yuan)
AI (5)	10000	50000	0.01	500
Student (7)	8000	56000	0.55	30800
Doctor(3)	2000	6000	2.2	13200
Expert (1) *	10000	10000	1.1	11000
Data Processing				10000
Organization				5000
			Grand Total (Yuan)	**70500**

* The expert round is only need to ensure >95% accuracy.

Future Directions

- Optimization based on individual AI translator accuracy
- Apply the crow-sourcing strategy to UMLS concept mapping
- Utilize the same strategy for translating English ontologies into Chinese and collect Chinese term variations for the same concept
- Apply our results to medical record mining and patient management

University of Michigan Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部 · PEKING UNIVERSITY HEALTH SCIENCE CENTER · 1912

Main Conclusions

- AI-based translators still do not provide satisfactory results for Chinese medical terms
- However, we can easily achieve 90-95% consistency by combining AI translation results from multiple sources and crowd sourcing confirmation by medical students and doctors.
- Our strategy is highly efficient and cost-effective.
- Once the thesaurus for a specific medical area is built, it can be used to mapping concepts in millions of Chinese medical records.

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912

Thanks and

Key Collaborators in this project:

Manhong Dai (MBNI, U of Michigan)

Huiying Qin (PKUHSC Natural Science in Medicine)

Qingbian Ma (PKUHSC Third Hospital)

Yanfang Wang (PKUHSC Clinical Research Institute)

University of Michigan
Health System

UMHS-PUHSC JOINT INSTITUTE

北京大学医学部
PEKING UNIVERSITY HEALTH SCIENCE CENTER
1912