Learning2extract for Medical Domain Retrieval

Yue Wang, Kuang Lu and Hui Fang

InfoLab

UNIVERSITY OF DELAWARE

Medical domain retrieval

- Finding literatures based on a given patient is important

Medical domain retrieval

- Query in medical domain tend to be complex

78 M w/ pmh of CABG in early [**Month (only) 3**] at [**Hospital64406**] (transferred to nursing home for rehab on [**12-8**] after several falls out of bed.) He was then readmitted to [**Hospital6 1749**] on [**3120-12-11**] after developing acute pulmonary edema/CHF/unresponsiveness?. The whether he had a small MI; he reportedly had a small NQWMI. He diuresis and was not intubated. . Yesterday, he was noted to have earlier this evening and then approximately 9 loose BM w/ some r frank blood just prior to transfer, unclear quantity.

Which terms should be used in the query

Key Terms

- The terms that could be helpful for retrieving relevant documents in medical domain.

78 M w/ pmh of CABG in early [**Month (only) 3**] at [**Hospital64406**] (transferred to nursing home for rehab on [**12-8**] after several falls out of bed.) He was then readmitted to [**Hospital6 1749**] on [**3120-12-11**] after developing acute pulmonary edema/CHF/unresponsiveness?. There was a question whether he had a small MI; he reportedly had a small NQWMI. He improved with diuresis and was not intubated. . Yesterday, he was noted to have a melanotic stool earlier this evening and then approximately 9 loose BM w/ some melena and some frank blood just prior to transfer, unclear quantity.

A Classification problem

Key term selection

Description query

78 M transferred to nursing home for rehab after CABG. Reportedly readmitted with a small NQWMI. Yesterday, he was noted to have a melanotic stool and then today he had approximately 9 loose BM w/ some melena and some frank blood just prior to transfer, unclear quantity.

Summary query

A 78 year old male presents with frequent stools and melena.

How?

Key term selection

Description query

78 M transferred to nursing home for rehab after CABG. Reportedly readmitted with a small NQWMI. Yesterday, he was noted to have a melanotic stool and then today he had approximately 9 loose BM w/ some melena and some frank blood just prior to transfer, unclear quantity.

Summary query

A 78 year old male presents with frequent stools and melena.

Domain features

Lexicon features

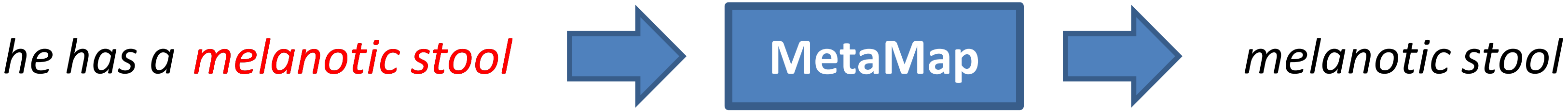
Statistical features

Locality features

POS features

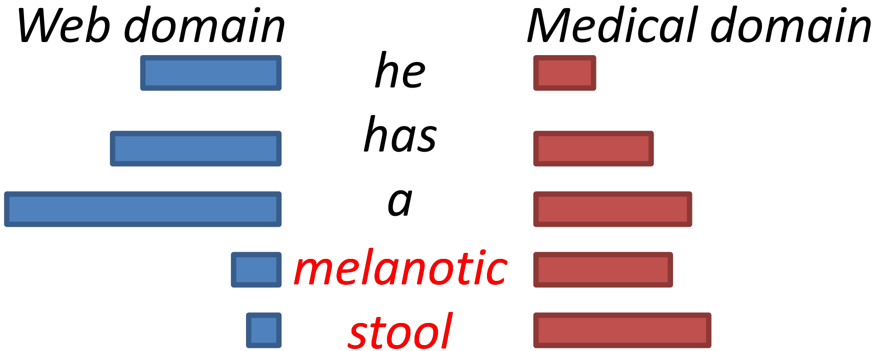
Domain features

- Concept(t_i)



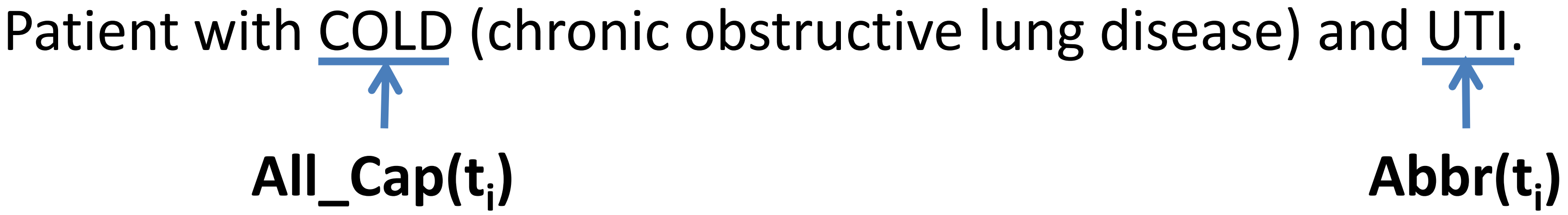
he has a melanotic stool → MetaMap → *melanotic stool*

- Unique(t_i)



Lexical features

- $\text{Abbr}(t_i)$
- $\text{All\_Cap}(t_i)$



Patient with COLD (chronic obstructive lung disease) and UTI.

$\text{All\_Cap}(t_i)$

$\text{Abbr}(t_i)$

Other features

Type	Feature	Description
Lexicon	Capitalized(t_i)	whether t_i contains any capital letters
	Stop(t_i)	whether t_i is a stopword
	Numeric(t_i)	whether t_i is a number
POS	Noun(t_i)	whether t_i is a noun, or part of a noun phrase
	Verb(t_i)	whether t_i is a verb, or part of a verb phrase
	Adj(t_i)	whether t_i is an adjective

Other features

Type	Feature	Description
Statistical	$tf_{des}(t_i)$	the term frequency in description of t_i
	$tf_c(t_i)$	the term frequency in collection of t_i
	$IDF(t_i)$	the invert document frequency t_i
	$wig(t_i)$	the weighted information gain of t_i
Locality	$Rank_{des}(t_i)$	the position of t_i shown in the description
	$Rank_{sent}(t_i)$	the position of sentence that contains t_i shown in the description

Key term identification results

	Precision	Recall	F1
Random Forest	0.753	0.631	0.686
Logistic Regression	**0.759**	0.636	**0.731**
Decision Tree	0.642	**0.821**	0.720
SVM	0.735	0.668	0.699

Logistic regression is used as the identification method

Applying key term for retrieval

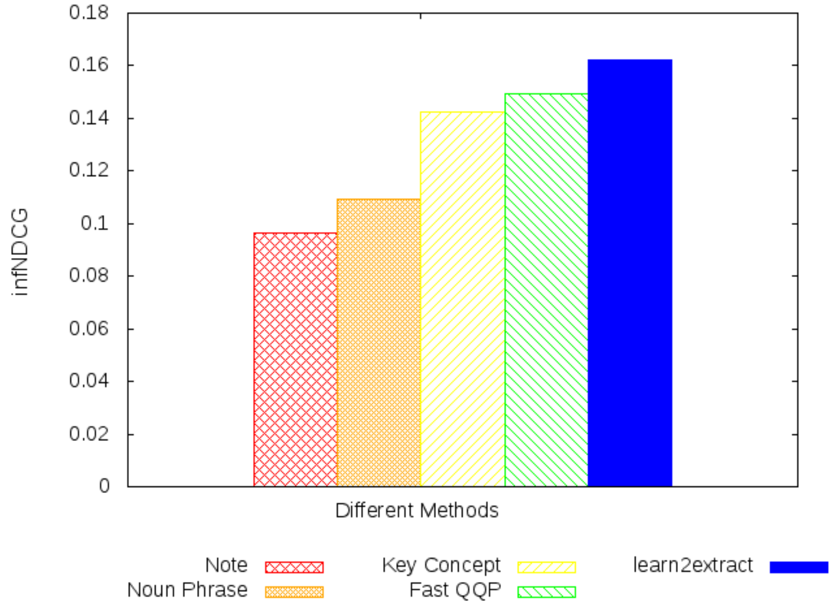
- Description query.

	CDS14	CDS15	CDS16
Summary (Upper Bound)	0.1712	0.2067	0.1844
Description	0.1397	0.1615	0.1537
Noun Phrase	0.1195	0.1487	0.1322
Key Concept	0.1426	0.1657	0.1594
Fast QQP	0.1498	0.1753	0.1584
Learn2extract	**0.1583**	**0.1779**	**0.1647**

Our method could outperform the baseline methods on all data collections

Applying key term for retrieval

- Note query



Best tuned performance drops

Proposed method still improve the performance over note query

Feature importance

- Domain features are the most helpful ones
- Followed by *abbr(t_i)*, *IDF(t_i)* and *wig(t_i)*

	Domain	Lexicon	POS	Statistical	Locality
CDS14	-0.067	-0.025	-0.005	-0.058	-0.004
CDS15	-0.074	-0.037	0.013	-0.047	0.003
CDS16	-0.066	-0.028	-0.004	-0.045	-0.007

Example of Identified Key Terms – description query

Summary	A 78 year old male presents with frequent stools and melena.
Noun Phrase	nursing home a small NQWMI a melanotic stool 9 loose BM some melena and some frank blood
Key Concept	nursing home CABG a small NQWMI noted stool prior to transfer
Fast QQP	nursing CABG readmitted with a small NQWMI melanotic stool approximately loose melena
learn2extract	rehab CABG NQWMI melanotic stool BM melena frank blood

Example of Identified Key Terms – Note query

Summary	A 78 year old male presents with frequent stools and melena.
Noun Phrase	CABG home acute pulmonary edema unresponsiveness a small MI NQWMI diuresis loose BM melanotic stool frank blood unclear quantity
Key Concept	CABG nursing home acute pulmonary edema CHF unresponsiveness small NQWMI melanotic stool loose BM
Fast QQP	pmh CABG nursing home falls bed pulmonary edema CHF unresponsiveness diuresis was not intubated melanotic loose frank blood
learn2extract	pmh CABG nursing home rehab pulmonary edema CHF NQWMI melanotic stool loose BM melena frank blood

Conclusion

- Key Term selection is important to improve the retrieval performance in medical domain
- A new set of features is proposed to identify Key Terms
- The retrieval performance could outperform the baseline methods using the selected features

Thank you!

Q&A