Cannabis_TREATS_cancer: Incorporating Fine-Grained ontological relations in Medical Document Retrieval

Author: Yunqing Xia, Zhongda Xie, Qiuge Zhang, Huiyuan Zhao, Huan Zhao
Presenter: Zhongda Xie

Outline

1. Introduction
2. Motivation
3. Methodology
4. Experiments
5. Conclusion
6. Future Work

1. Introduction(1/3)

- Traditional IR systems are based on independent keywords, which are called bag-of-word models.

- The ignorance of the connection between the words may lead to mistakes:
  - Query: Cannabis and Cancer
  - (Sentence-one) He is in bad conditions, he suffers from *cancer*, and he's addicted to *cannabis*.
  - (Sentence-two) Studies prove that *cannabis* can be an effective treatment for *cancer*.[TREATS]
  - (Sentence-three)The report indicates that long-term *cannabis* use may cause lung *cancer*.[CAUSES]

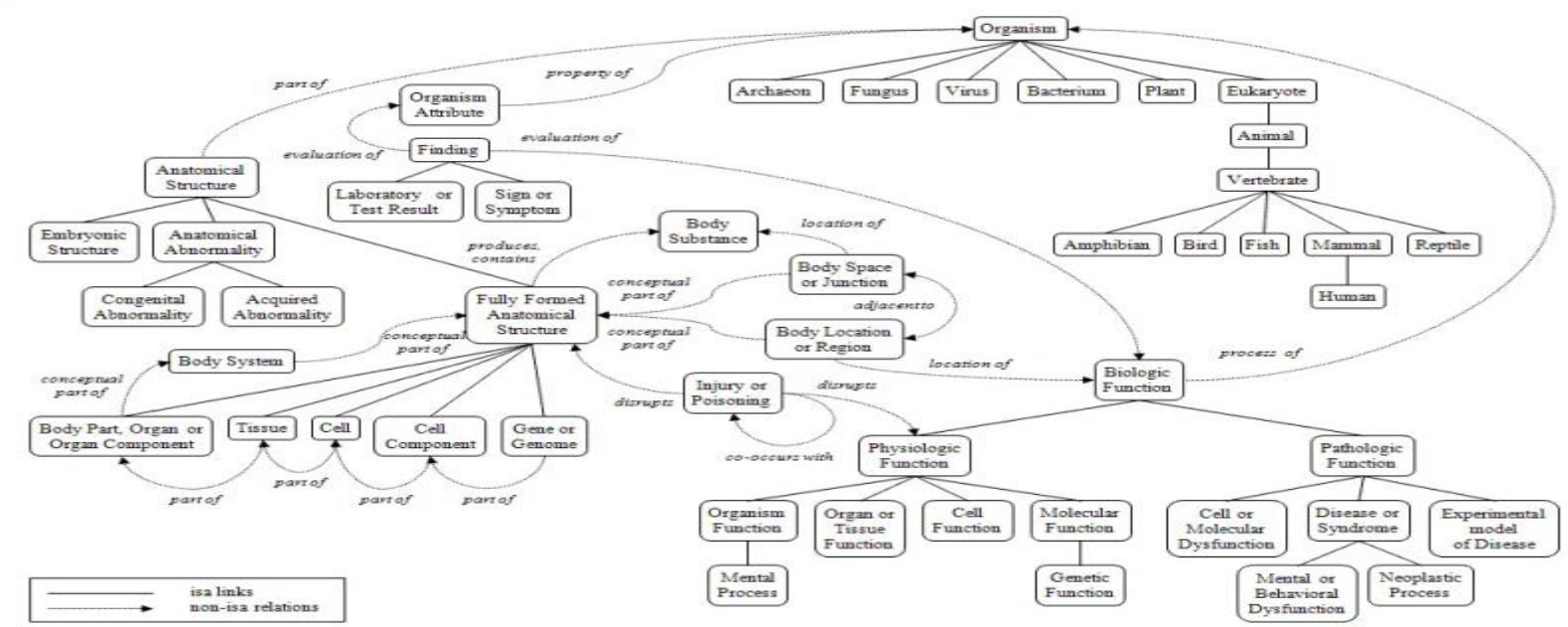
1. Introduction(2/3)

- Previous work has justified the assumption that relations of various linguistic levels are helpful to improve document ranking:
  - Statistical term dependency:
    - Gao [4] linkage dependency language model
    - Hou[8] Higher-Order word association relation
  - Coarse-grained relations:
    - Park[7] quasi-synchronous dependence model
    - Lu[11] structural representation of texts
    - Khoo[12] cause-effect relation
    - Li[13] semantic relations (but are too general)

- The Coarse-grained relations are too general(like *is_a*, *co-occurs_with*).
- Fine-grained relations have real meainings(like *treat*, *is_symptom_of*, *diagnoses*).

1. Introduction(3/3)

- In specific areas like Medical Information, there are some ontologies which have fine-grained relations.
- SemMedDB[18]



- Vintar[3] achieved positive results, but they use it to filter cross-lingual web pages in a boolean manner.

2. Motivation

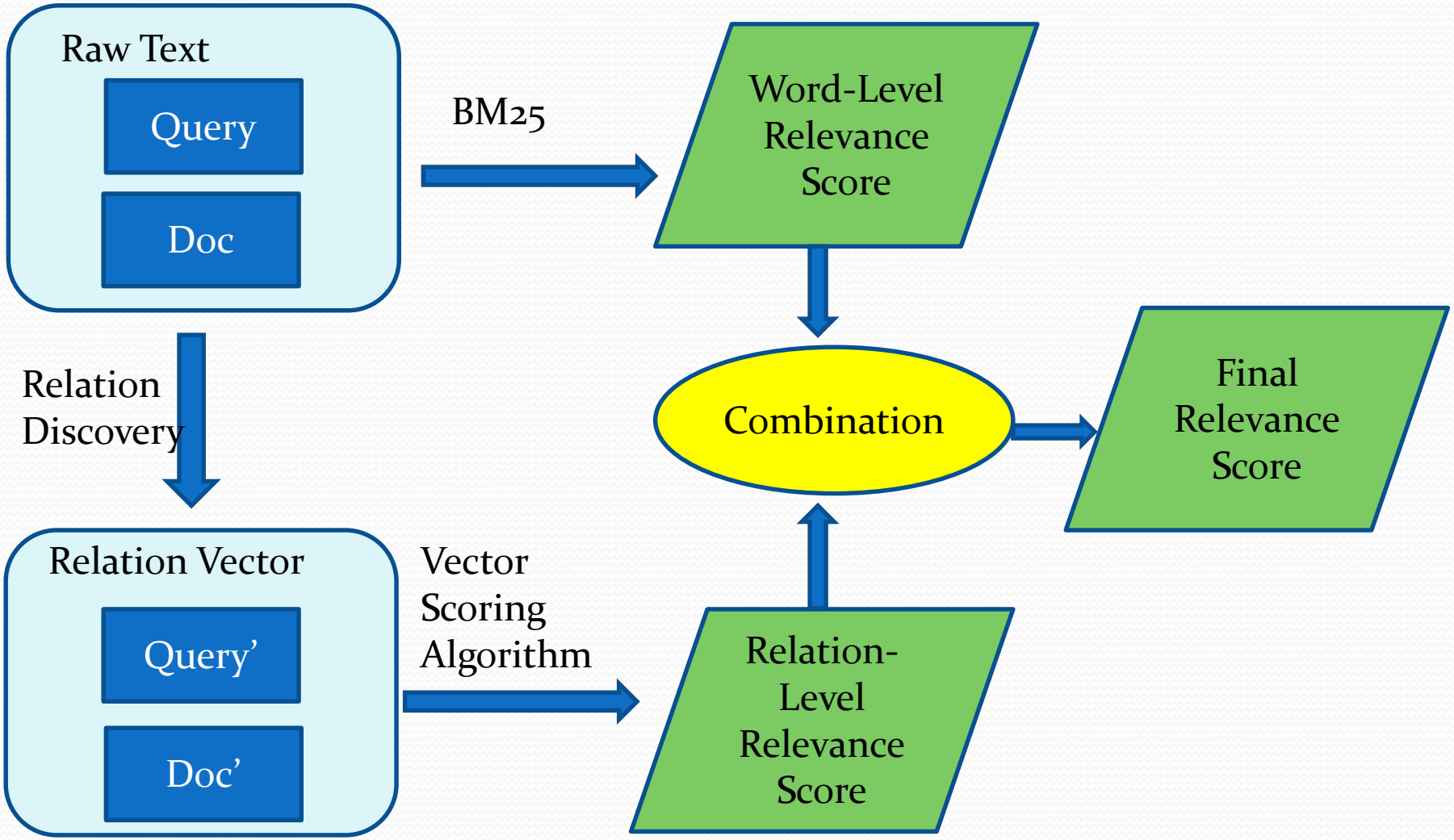
- Use fine-grained ontological relation(SemMedDB) in specific area(Medical) Document Retrieval, and to find whether it can improve retrieval result.

- To propose an algorithm to evaluate the query-document relevance score in relation level.

- To determine a better way of combining the relevance scores in relation-level and traditional word-level

3. Methodology

Framework



3. Methodology

- Ontological Relation Detection

- SemMedDB has 57 kinds of ontological relations, and we choose 18 of them:
  - PROCESS OF, METHOD OF, LOCATION OF, PART OF, OCCURS IN, STIMULATES, MANIFESTATION OF, CONVERT TO, AUGMENTS, ASSOCIATED WITH, PREVENTS, USES, TREATS, PREDISPOSES, PRODUCES, DISRUPTS, CAUSES and INHIBITS

- Use the predicate instances of the relation for detection.
  - Studies prove that cannabis can be an effective *treatment* for cancer.
  - The report show that his cancer may be *treated* by the right amount of cannabis.

3. Methodology

- Representation of Query and Document Using Ontological Relation

- Query
  - Queries are often too short to detect any relation keyword

  - Cannabis and Cancer
  - (0,0,0,0,0.5,0,0,0,0.5,0,0,0,0,0,0,0,0,0,0) [TREATS, CAUSES]

- Document

  (1)He is in bad conditions, he suffers from cancer, and he's addicted to cannabis. (2)His cough is treated by using Aspirin. (3)The report indicates that long-term cannabis use may cause lung cancer. (4)Studies prove that cannabis can be an effective treatment for cancer. (5)The report show that his cancer may be treated by the right amount of cannabis.

  (0,0,0,0,2/3,0,0,0,1/3,0,0,0,0,0,0,0,0,0,0)

3. Methodology

- Relation Relevance Score: Cosine Distance

$$Cos(v_i, v_j) = \frac{\sum_{k=1}^{18} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{18} w_{ik}^2 + \sum_{k=1}^{18} w_{jk}^2}}$$

- Combination Method:
  - r: word-level relevance score;
  - l: relation-level relevance score.

I. Summation

$$r^* = \alpha \times r + (1 - \alpha) \times l$$

II. Multiplication

$$r^* = r \times l$$

III. Amplification

$$r^* = r \times \beta^l$$

4. Experiments

Data Set:

- CLEF: Clef2013 eHealth Lab Medical IR task[17]
- CLEF+: Extended non-annotation documents
  (three medical students as assessors, Kappa co-efficient 0.82)

Queries:

- 14 out of all 50 queries contain two concepts.

Evaluation Metrics:

① p@10: precision at top 10.
② nDCG@10: normalized Discounted Cumulative Gain
③ MAP: Mean average precision at top 10.

4. Experiments

Relation Detection Window

① **CURS**: The current sentence
② **CURSP**: The current and the preceding sentence
③ **CURSPF**: The current, preceding, following sentences
④ **CURP**: The current paragraph
⑤ **CURD**: The current web document
⑥ **HTML**: Text in the current HTML tag pair relation

Method	p@10		nDCG@10		MAP@10	
	CLEF	CLEF+	CLEF	CLEF+	CLEF	CLEF+
CURS	0.450	0.522	0.445	0.511	0.110	0.131
CURSP	0.451	0.524	0.448	0.513	0.111	0.134
CURSPF	0.453	0.528	0.449	0.516	0.112	0.137
CURP	0.442	0.476	0.431	0.502	0.106	0.127
CURD	0.427	0.458	0.416	0.489	0.097	0.112
HTML	0.456	0.534	0.452	0.519	0.124	0.143

Conclusion: HTML is the best

4. Experiments

Combination Method:

① SUUM: the summation method, α empirically set 0.7

② MULT: the multiplication method

③ AMPL: the amplification method.

Method	p@10		nDCG@10		MAP@10	
	CLEF	CLEF+	CLEF	CLEF+	CLEF	CLEF+
SUMM	0.451	0.527	0.447	0.511	0.121	0.140
MULT	0.447	0.521	0.443	0.501	0.117	0.139
AMPL	0.456	0.534	0.452	0.519	0.124	0.144

AMPL outperforms the other two over all metrics.

4. Experiments

Different Methods:

① BM25: Okapi BM25. Default setting.

② BMB: the method in [3], filtering web pages in a boolean manner.

③ BMR: Our method, combing word-level and relation-level score.

Method	p@10		nDCG@10		MAP@10	
	CLEF	CLEF+	CLEF	CLEF+	CLEF	CLEF+
BM25	0.450	0.516	0.448	0.504	0.112	0.129
BMB	0.437	0.521	0.435	0.514	0.106	0.117
BMR	0.456	0.534	0.452	0.519	0.124	0.144

5. Conclusion

1. We propose a novel medical document ranking method, which incorporates the fine-grained ontological relations in relevance scoring.
2. We think of a way to evaluate the relation-level relevance of query and document.
3. We explore the influence of combination model and relation detection window.
4. We compared the result with some related works, and it turns out better.

6. Future Work

1. The 18 relations are compiled by human experts, and we hope to extend these relations to cover all the possible relations.

2. To propose a better relation detection algorithm.

3. Apply ontological relation method in general domain.

4. Conduct more experiments, comparing with other methods.

References

1. Song, F., Croft, W.B.: A general language model for information retrieval. In: Proc. of CIKM 1999, pp. 316–321. ACM, New York (1999)

2. Matsumura, A., Takasu, A.: Adachi: The effect of information retrieval method using dependency relationship between words. In: Proceedings of RIAO 2000, pp. 1043–1058 (2000)

3. Vintar, S., Buitelaar, P., Volk, M.: Semantic relations in concept-based crosslanguage medical information retrieval. In: Proceedings of ECML/PKDD workshop on Adaptive Text Eextraction and Mining (ATEM) (2003)

4. Gao, J., Nie, J.Y., Wu, G., Cao, G.: Dependence language model for information retrieval. In: Proc. of SIGIR 2004, pp. 170–177. ACM, New York (2004)

5. Morton, T.: Using semantic relations to improve information retrieval. PhD thesis, University of Pennsylvania (2004)

6. Maisonnasse, L., Gaussier, E., Chevallet, J.P.: Revisiting the dependence language model for information retrieval. In: Proc. of SIGIR 2007, pp. 695–696. ACM, New York (2007)

7. Park, J.H., Croft, W.B., Smith, D.A.: A quasi-synchronous dependence model for information retrieval. In: Proc. of CIKM 2011, pp. 17–26. ACM, New York (2011)

8. Hou, Y., Zhao, X., Song, D., Li, W.: Mining pure high-order word associations via information geometry for information retrieval. ACM Trans. Inf. Syst. 31(3), 12:1–12:32 (2013)

9. Zhao, J., Huang, J.X., Ye, Z.: Modeling term associations for probabilistic information retrieval. ACM Trans. Inf. Syst. 32(2), 7:1–7:47 (2014)

References

10. Giger, H.P.: Concept based retrieval in classical ir systems. In: Proc. of SIGIR 1988, pp. 275–289. ACM, New York (1988)

11. Lu, X.: Document retrieval: A structural approach. Inf. Process. Manage. 26(2), 209–218 (1990)

12. Khoo, C.S.G., Myaeng, S.H., Oddy, R.N.: Using cause-effect relations in text to improve information retrieval precision. Inf. Process. Manage. 37(1), 119–145 (2001)

13. Li, Y., Wang, Y., Huang, X.: A relation-based search engine in semantic web. IEEE Trans. on Knowl. and Data Eng. 19(2), 273–282 (2007)

14. Lee, J., Min, J.K., Oh, A., Chung, C.W.: Effective ranking and search techniques for web resources considering semantic relationships. Inf. Process. Manage. 50(1), 132–155 (2014)

15. Bilotti, M.W., Elsas, J., Carbonell, J., Nyberg, E.: Rank learning for factoid question answering with linguistic and semantic constraints. In: Proc. of CIKM 2010,pp. 459–468. ACM, New York (2010)

16. Voorhees, E.M., Hersh, W.: Overview of the trec 2012 medical records track. In: Proc. of TREC 2012 (2012)

17. Goeuriot, L., Jones, G.J.F., Kelly, L., Leveling, J., Hanbury, A., M ¨ uller, H., Salanter ¨ a, S., Suominen, H., Zuccon, G.: Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In: CLEF Online Working Notes (2013)

18. Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., Rindflesch, T.C.: Semmeddb: a pubmed-scale repository of biomedical semantic predications. Bioinformatics 28(23), 3158–3160 (2012)

Thank you!