Keyphrases as Knowledge Units for Text-based Applications

Daqing He

Department of Informatics and Networked Systems

School of Computing and Information

University of Pittsburgh

Outline

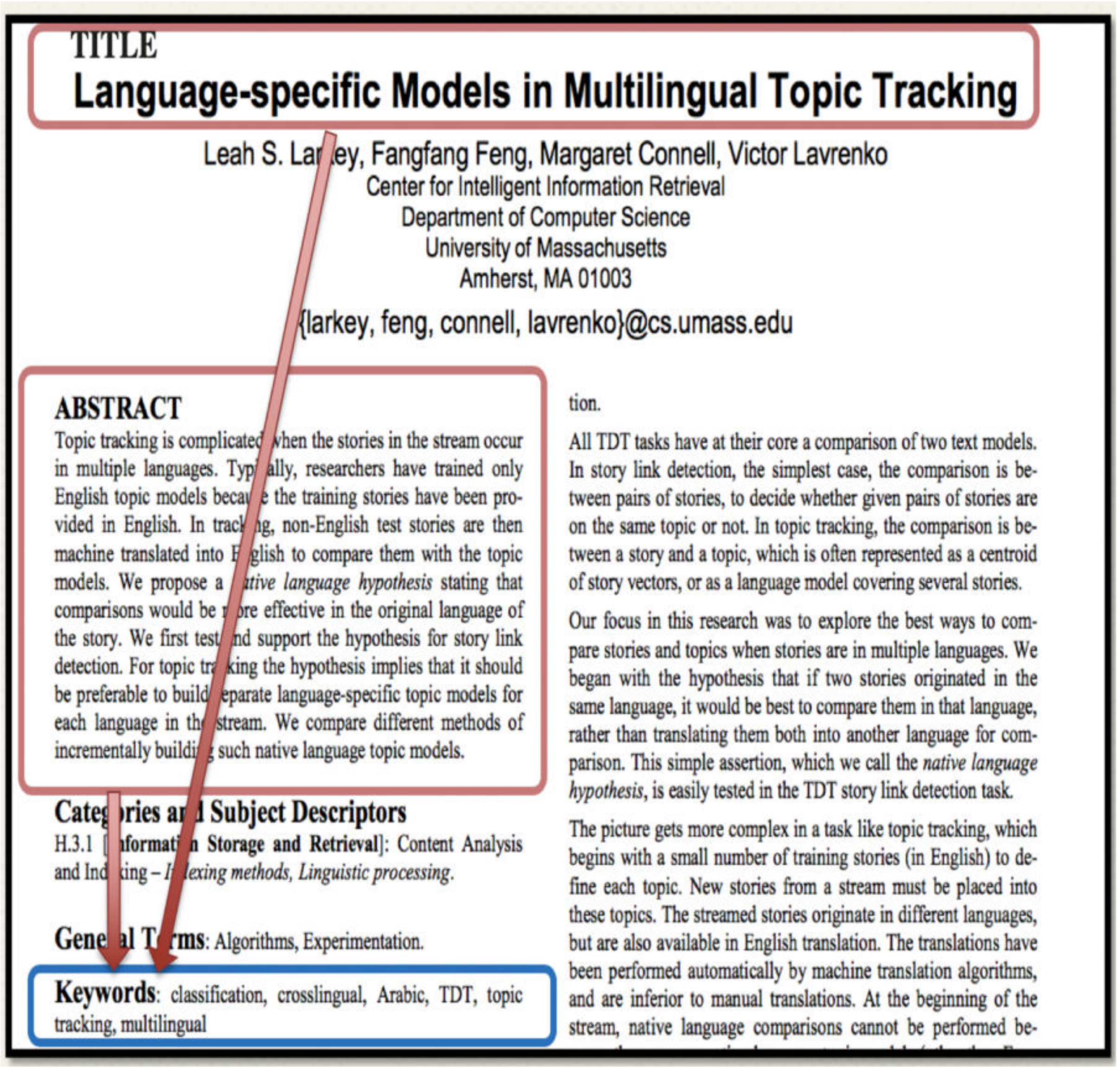
- Basics of Keyphrases: Definitions and Importance
- Identification of Keyphrases: Extraction and Generation
- Applications of Keyphrases: knowledge unit for supporting student learning
- Applications of Keyphrases: knowledge unit for recognizing patients' concerns
- Applications of Keyphrases: knowledge unit for interactive machine learning
- Conclusions

Outline

- Basics of Keyphrases: Definitions and Importance
- Identification of Keyphrases: Extraction and Generation
- Applications of Keyphrases: knowledge unit for supporting student learning
- Applications of Keyphrases: knowledge unit for recognizing patients’ concerns
- Applications of Keyphrases: knowledge unit for interactive machine learning
- Conclusions

Keyphrases: a Definition

- Keyphrases: Short noun phrases to summarize and highlight important information in a piece of text
  - Examples: see the figure



TITLE

Language-specific Models in Multilingual Topic Tracking

Leah S. Larkey, Fangfang Feng, Margaret Connell, Victor Lavrenko
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{larkey, feng, connell, lavrenko}@cs.umass.edu

ABSTRACT

Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare them with the topic models. We propose a *native language hypothesis* stating that comparisons would be more effective in the original language of the story. We first test and support the hypothesis for story link detection. For topic tracking the hypothesis implies that it should be preferable to build separate language-specific topic models for each language in the stream. We compare different methods of incrementally building such native language topic models.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Indexing methods, Linguistic processing.*

General Terms: Algorithms, Experimentation.

Keywords: classification, crosslingual, Arabic, TDT, topic tracking, multilingual

tion.

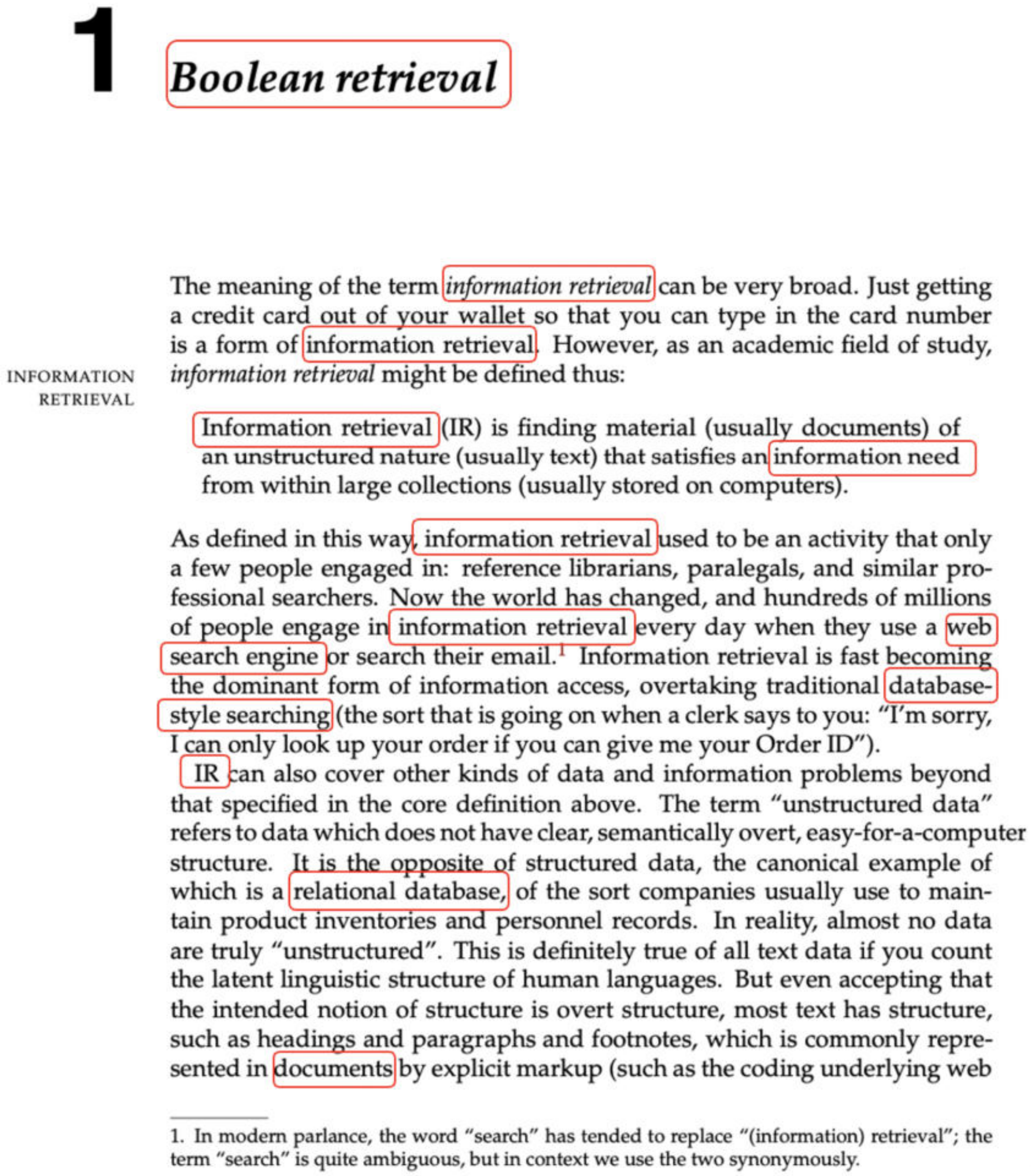
All TDT tasks have at their core a comparison of two text models. In story link detection, the simplest case, the comparison is between pairs of stories, to decide whether given pairs of stories are on the same topic or not. In topic tracking, the comparison is between a story and a topic, which is often represented as a centroid of story vectors, or as a language model covering several stories.

Our focus in this research was to explore the best ways to compare stories and topics when stories are in multiple languages. We began with the hypothesis that if two stories originated in the same language, it would be best to compare them in that language, rather than translating them both into another language for comparison. This simple assertion, which we call the *native language hypothesis*, is easily tested in the TDT story link detection task.

The picture gets more complex in a task like topic tracking, which begins with a small number of training stories (in English) to define each topic. New stories from a stream must be placed into these topics. The streamed stories originate in different languages, but are also available in English translation. The translations have been performed automatically by machine translation algorithms, and are inferior to manual translations. At the beginning of the stream, native language comparisons cannot be performed be-

Keyphrases: a Definition

- Keyphrases: Short noun phrases to summarize and highlight important information in a piece of text
  - Examples: see the figure



1 Boolean retrieval

The meaning of the term *information retrieval* can be very broad. Just getting a credit card out of your wallet so that you can type in the card number is a form of information retrieval. However, as an academic field of study, *information retrieval* might be defined thus:

INFORMATION RETRIEVAL

> Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

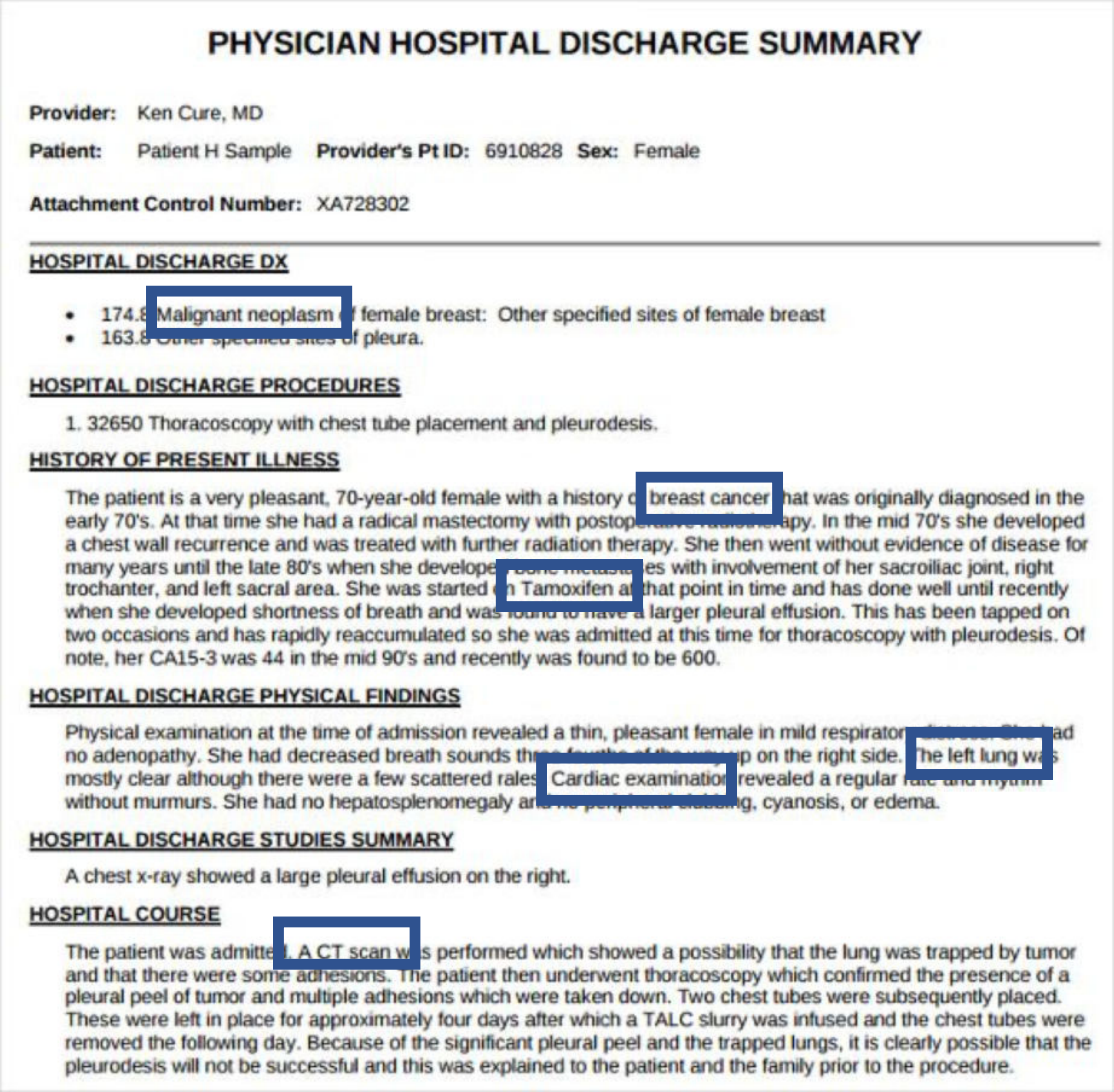
As defined in this way, information retrieval used to be an activity that only a few people engaged in: reference librarians, paralegals, and similar professional searchers. Now the world has changed, and hundreds of millions of people engage in information retrieval every day when they use a web search engine or search their email.[1] Information retrieval is fast becoming the dominant form of information access, overtaking traditional database-style searching (the sort that is going on when a clerk says to you: "I'm sorry, I can only look up your order if you can give me your Order ID").

IR can also cover other kinds of data and information problems beyond that specified in the core definition above. The term "unstructured data" refers to data which does not have clear, semantically overt, easy-for-a-computer structure. It is the opposite of structured data, the canonical example of which is a relational database, of the sort companies usually use to maintain product inventories and personnel records. In reality, almost no data are truly "unstructured". This is definitely true of all text data if you count the latent linguistic structure of human languages. But even accepting that the intended notion of structure is overt structure, most text has structure, such as headings and paragraphs and footnotes, which is commonly represented in documents by explicit markup (such as the coding underlying web

1. In modern parlance, the word "search" has tended to replace "(information) retrieval"; the term "search" is quite ambiguous, but in context we use the two synonymously.

Keyphrases: a Definition

- Keyphrases: Short noun phrases to summarize and highlight important information in a piece of text
  - Examples: see the figure



PHYSICIAN HOSPITAL DISCHARGE SUMMARY

Provider: Ken Cure, MD

Patient: Patient H Sample **Provider's Pt ID:** 6910828 **Sex:** Female

Attachment Control Number: XA728302

HOSPITAL DISCHARGE DX

- 174.8 Malignant neoplasm of female breast: Other specified sites of female breast
- 163.8 Other specified sites of pleura.

HOSPITAL DISCHARGE PROCEDURES

1. 32650 Thoracoscopy with chest tube placement and pleurodesis.

HISTORY OF PRESENT ILLNESS

The patient is a very pleasant, 70-year-old female with a history of breast cancer that was originally diagnosed in the early 70's. At that time she had a radical mastectomy with postoperative radiation therapy. In the mid 70's she developed a chest wall recurrence and was treated with further radiation therapy. She then went without evidence of disease for many years until the late 80's when she developed bone metastases with involvement of her sacroiliac joint, right trochanter, and left sacral area. She was started on Tamoxifen at that point in time and has done well until recently when she developed shortness of breath and was found to have a larger pleural effusion. This has been tapped on two occasions and has rapidly reaccumulated so she was admitted at this time for thoracoscopy with pleurodesis. Of note, her CA15-3 was 44 in the mid 90's and recently was found to be 600.

HOSPITAL DISCHARGE PHYSICAL FINDINGS

Physical examination at the time of admission revealed a thin, pleasant female in mild respiratory distress. She had no adenopathy. She had decreased breath sounds three-fourths of the way up on the right side. The left lung was mostly clear although there were a few scattered rales. Cardiac examination revealed a regular rate and rhythm without murmurs. She had no hepatosplenomegaly and no peripheral clubbing, cyanosis, or edema.

HOSPITAL DISCHARGE STUDIES SUMMARY

A chest x-ray showed a large pleural effusion on the right.

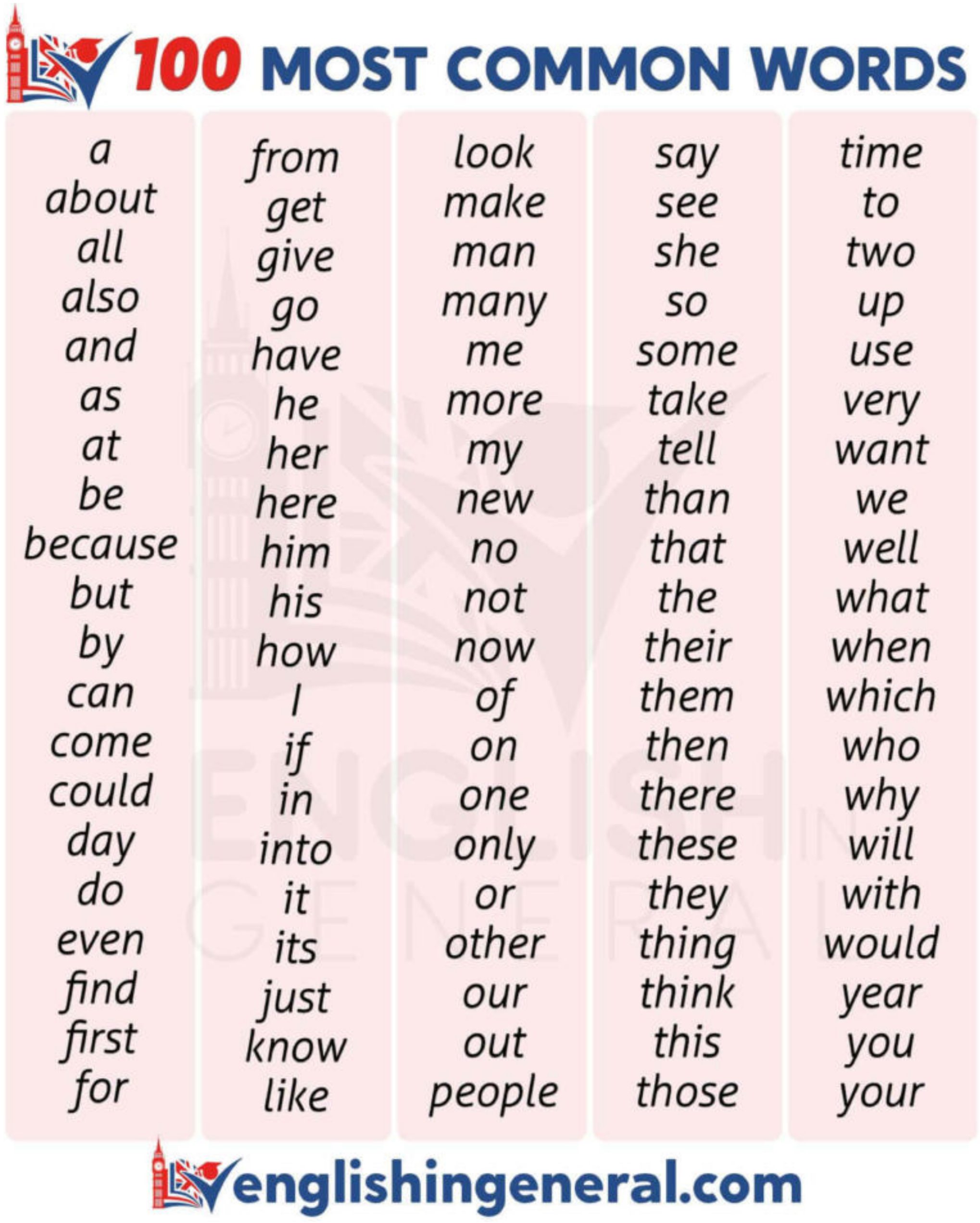
HOSPITAL COURSE

The patient was admitted. A CT scan was performed which showed a possibility that the lung was trapped by tumor and that there were some adhesions. The patient then underwent thoracoscopy which confirmed the presence of a pleural peel of tumor and multiple adhesions which were taken down. Two chest tubes were subsequently placed. These were left in place for approximately four days after which a TALC slurry was infused and the chest tubes were removed the following day. Because of the significant pleural peel and the trapped lungs, it is clearly possible that the pleurodesis will not be successful and this was explained to the patient and the family prior to the procedure.

Keyphrases: a Definition

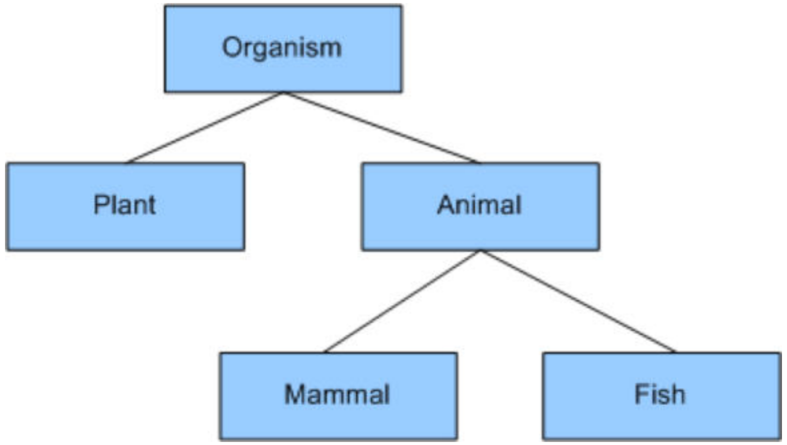
▷ Keyphrases: Short noun phrases to summarize and highlight important information in a piece of text
  - Examples: see the figure

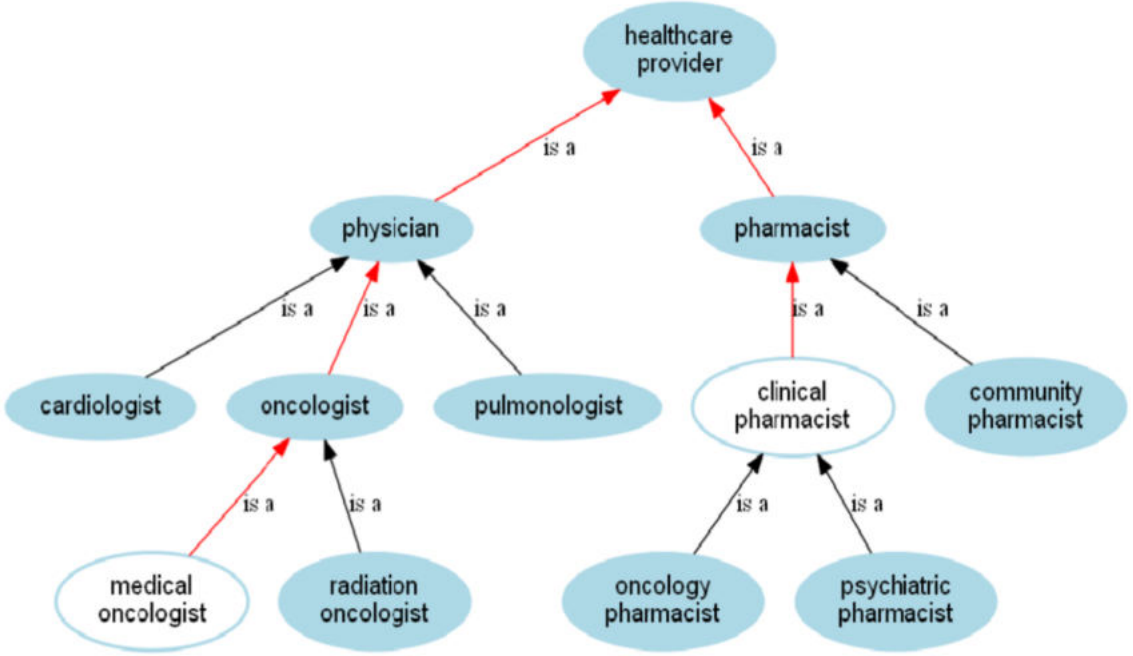
▷ Keyphrases are different to words
  - Same as keyphrases, string of characters written by authors
  - Different to keyphrases, individual words might not have a complete and unique meaning



Keyphrases: a Definition

- Keyphrases: Short noun phrases to summarize and highlight important information in a piece of text
  - Examples: see the figure
- Keyphrases are different to words
  - Same as keyphrases, string of characters written by authors
  - Different to keyphrases, individual words might not have a complete and unique meaning
- Keyphrases are different to concepts
  - Same as keyphrases, concepts have a complete and unique meaning, good knowledge unit
  - Different to keyphrases, concepts are more knowledge focus, less language focus, thus need manual construct



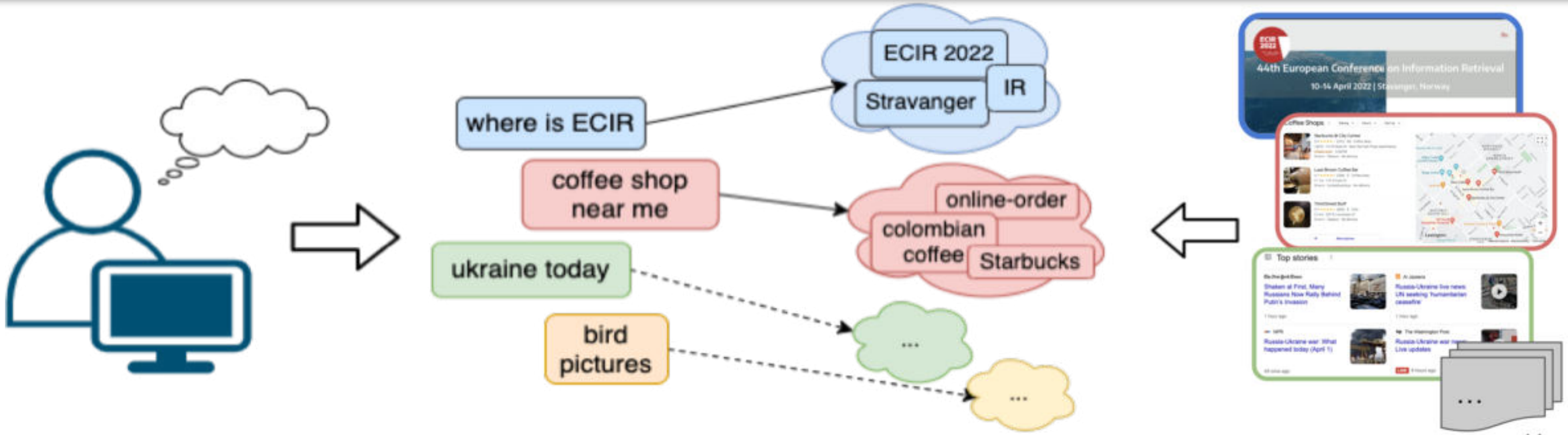


Why Keyphrases?

- ▷ Keyphrases are a natural and neat way to express important information ➔ semantic and knowledge unit
  - Better than words
- ▷ Keyphrases are a natural & efficient language units connecting human and data/information
  - Better than concepts, which need abstraction
- ▷ Modern representations of keyphrases enable wider applications of keyphrases
  - Embedding-based representations enable direct computation on keyphrases
  - Enable keyphrases act as knowledge unit, not just text level

Takeaway message

- Keyphrases can combine the benefits of words and concepts in various real world applications



Applications of Keyphrases

▷ Information retrieval (indexing term)

Jacquemin C. and Bourigault D. (2003). Term extraction and automatic indexing. In Mitkov R. (ed), The Oxford Handbook of Computational Linguistics, Oxford University Press, pp. 599–615.

Applications of Keyphrases

- Information retrieval (indexing term)
- Summarization (locate key sentences)

D'Avanzo, Ernesto, Bernardo Magnini, and Alessandro Vallin. "Keyphrase extraction for summarization purposes: The LAKE system at DUC-2004." Proceedings of the 2004 document understanding conference. 2004.

Applications of Keyphrases

- ▷ Information retrieval (indexing term)
- ▷ Summarization (locate key sentences)

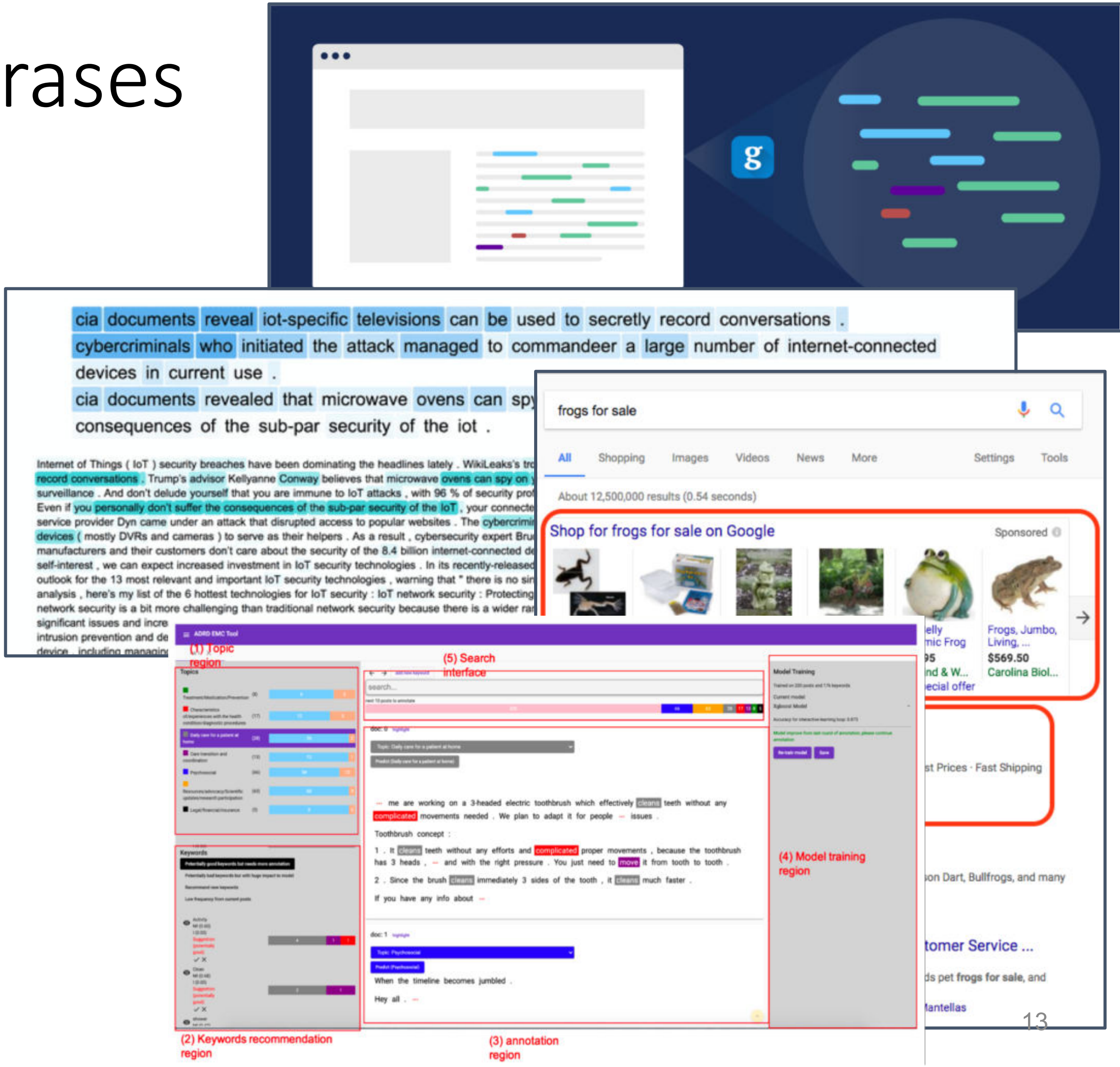
- ▷ Online Advertising

Yih, Wen-tau, Joshua Goodman, and Vitor R. Carvalho. "Finding advertising keywords on web pages." Proceedings of the 15th international conference on World Wide Web. 2006.

Applications of Keyphrases

- ▷ Information retrieval (indexing term)
- ▷ Summarization (locate key sentences)

- ▷ Online Advertising
- ▷ Many other applications

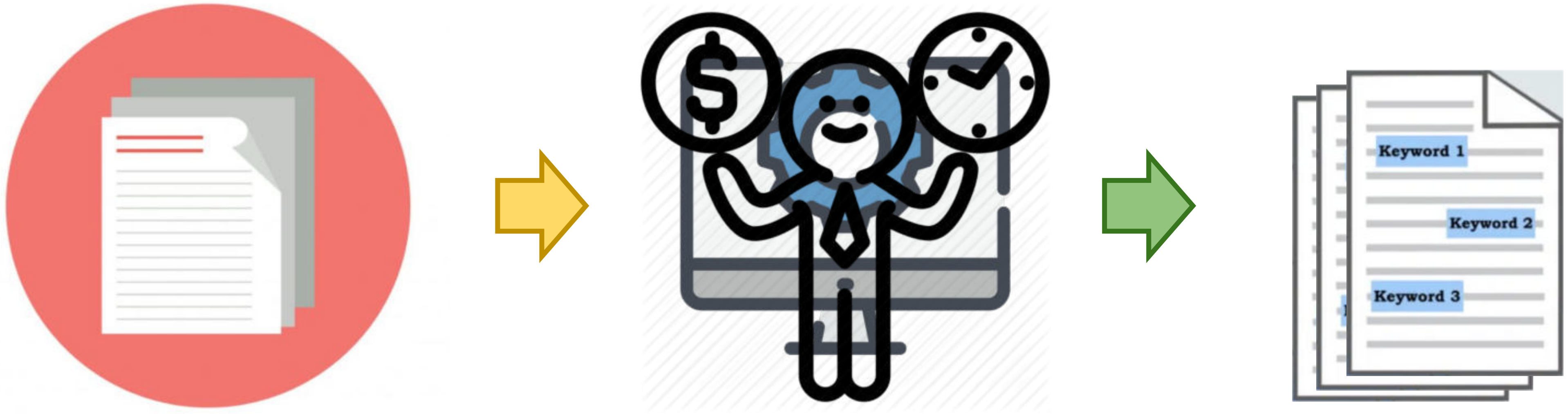


Outline

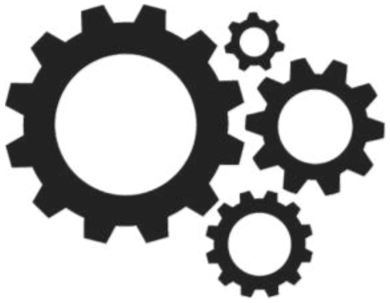
- Basics of Keyphrases: Definitions and Importance
- Identification of Keyphrases: Extraction and Generation
- Applications of Keyphrases: knowledge unit for supporting student learning
- Applications of Keyphrases: knowledge unit for recognizing patients’ concerns
- Applications of Keyphrases: knowledge unit for interactive machine learning
- Conclusions

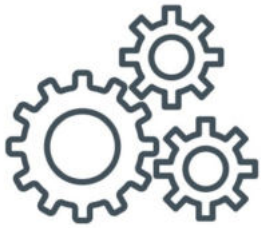
How to Obtain Keyphrases?

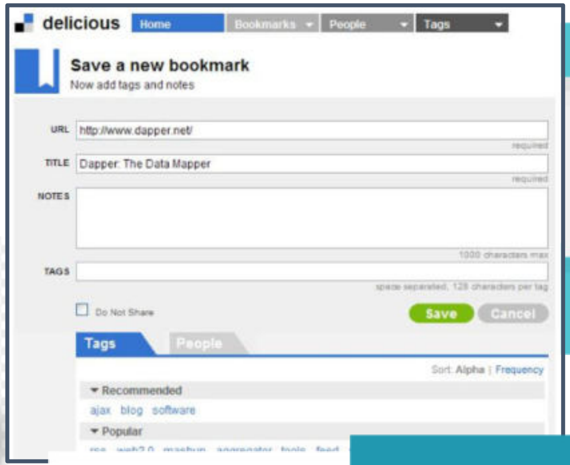
- ▷ Certainly not manual methods ==> avoid same limitation of concepts
- ▷ Ideal approaches should be automatic

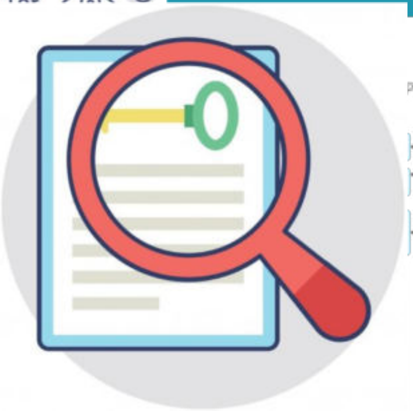


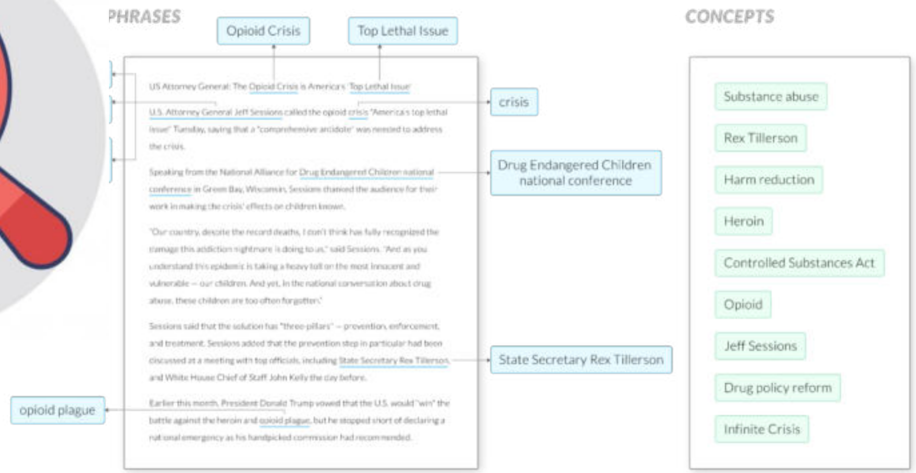
Digital Texts do Help!







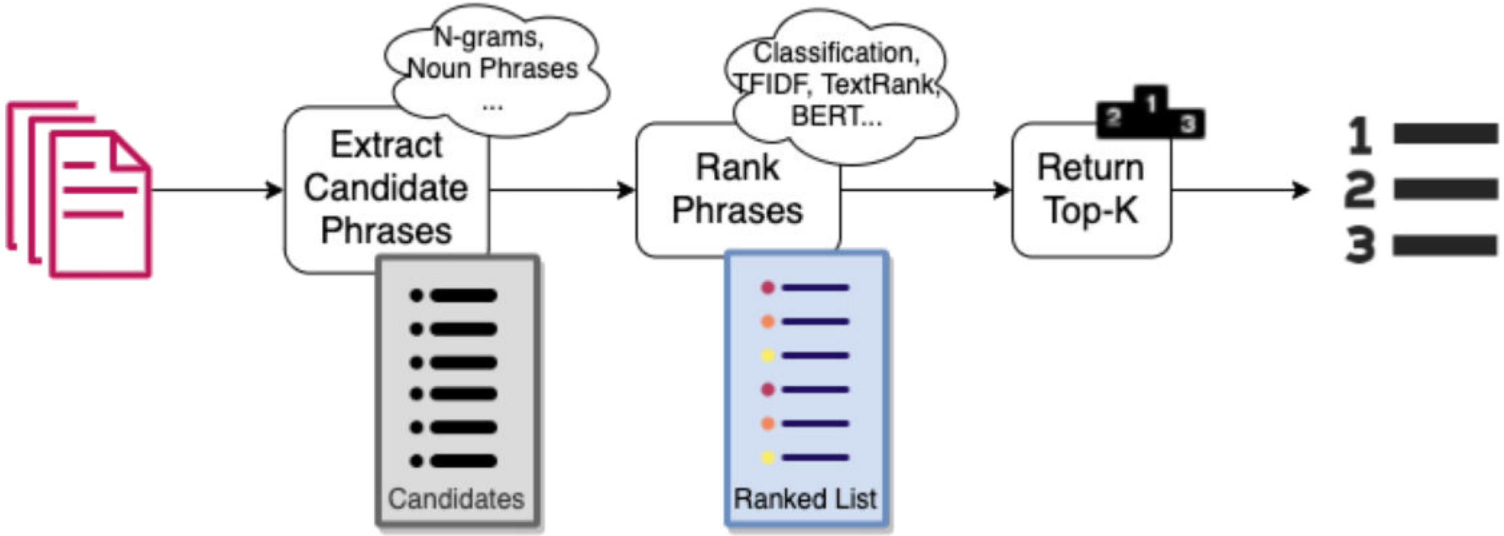




https://cs.stanford.edu/people/ebrun/papers/Keyphrase_for_education.pdf

Method 1: Keyphrase Extraction

- Select important words/phrases from the source text
  - Step 1: Generate candidates
  - Step 2: Rank candidates and return top K as results



Papagiannopoulou, Eirini, and Grigorios Tsoumakas. "A review of keyphrase extraction." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10.2 (2020): e1339.

Automatic Keyphrase Extraction in Textbooks

▷ Framework for manual concept annotation

INFORMATION RETRIEVAL

The meaning of the term *information retrieval* can be very broad. Just getting a credit card out of your wallet so that you can type in the card number is a form of information retrieval. However, as an academic field of study, *information retrieval* might be defined thus:

> Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

As defined in this way, information retrieval used to be an activity that only a few people engaged in: reference librarians, paralegals, and similar professional searchers. Now the world has changed, and hundreds of millions of people engage in information retrieval every day when they use a web search engine or search their email.[1] Information retrieval is fast becoming the dominant form of information access, overtaking traditional database-style searching (the sort that is going on when a clerk says to you: "I'm sorry, I can only look up your order if you can give me your Order ID").

IR can also cover other kinds of data and information problems beyond that specified in the core definition above. The term "unstructured data" refers to data which does not have clear, semantically overt, easy-for-a-computer structure. It is the opposite of structured data, the canonical example of which is a relational database, of the sort companies usually use to maintain product inventories and personnel records. In reality, almost no data are truly "unstructured". This is definitely true of all text data if you count the latent linguistic structure of human languages. But even accepting that the intended notion of structure is overt structure, most text has structure, such as headings and paragraphs and footnotes, which is commonly represented in documents by explicit markup (such as the coding underlying web

Table 1 Coding schema for concept annotation

Rule	Description (bold text) with Examples and Explanations
1. (Round 1)	**Only noun/noun phrases are considered** Examples: Concept: sorting algorithm, wildcard pattern matching, boolean retrieval model Not concept: merging postings list, ranking documents In the examples above, *merging postings list* and *ranking documents* are not concepts, because they are not nouns or noun phrases
2. (Round 1)	**Abbreviation of a concept is also a concept** Examples: "IR (information retrieval)", "EM (expectation maximization)" *IR* and *EM* are all concepts, because *information retrieval* and *expectation maximization* are concepts
3. (Round 1)	**Annotate the whole noun/noun phrases, but ignore the general adj. (e.g., long, big etc.)** Examples: Concept: latent linguistic structure, hidden variables Not concept: long query, big document collection In the examples above, long and big are too general. Only *query* and *document collection* are concepts
4. (Round 2)	**If two noun phrases are concepts, the combination should be the concept** Examples: Concept: postings list, data structure, postings list data structure In the example above, *postings list* and *data structure* are concepts, so *postings list data structure* is a concept.
5. (Round 3)	**The concepts combined with conjunctions should be separated (e.g., and, or)** Examples: "boolean and proximity queries" In the example above, you need to annotate the two concepts *boolean queries* and *proximity queries*
6. (Round 5)	**All variations of the concepts should be annotated** Examples: Concept: Multi-term query, Bi-term query, Three-term query The examples above are variations of the concept query, therefore they should be annotated.
7. (Round 6)	**Annotate all special / not general phrases in the Computer Science related domain e.g., Statistics, mathematics** Examples: Concept: quadratic function, binomial distribution *Quadratic function* and *binomial distribution* are concepts, because they are important phrases in the Statistics domain.
8. (Round 6)	**Ignore the Abbreviation in brackets** Examples: "inverse document frequency (idf)" "variable byte (vb)" "encodingmegabytes (mb)" In the examples above, idf, vb, and mb should be ignored
9. (Round 8)	**If the concept term has punctuations, keep them** Examples: "(query, document) pairs" The example above should be annotated as a concept including the bracket and comma.
10. (Round 9)	**The well-known and important examples should be annotated** Examples: "A well-known example is the Unified Medical Language System..." In the example above, *Unified Medical Language System* should be annotated as a concept.

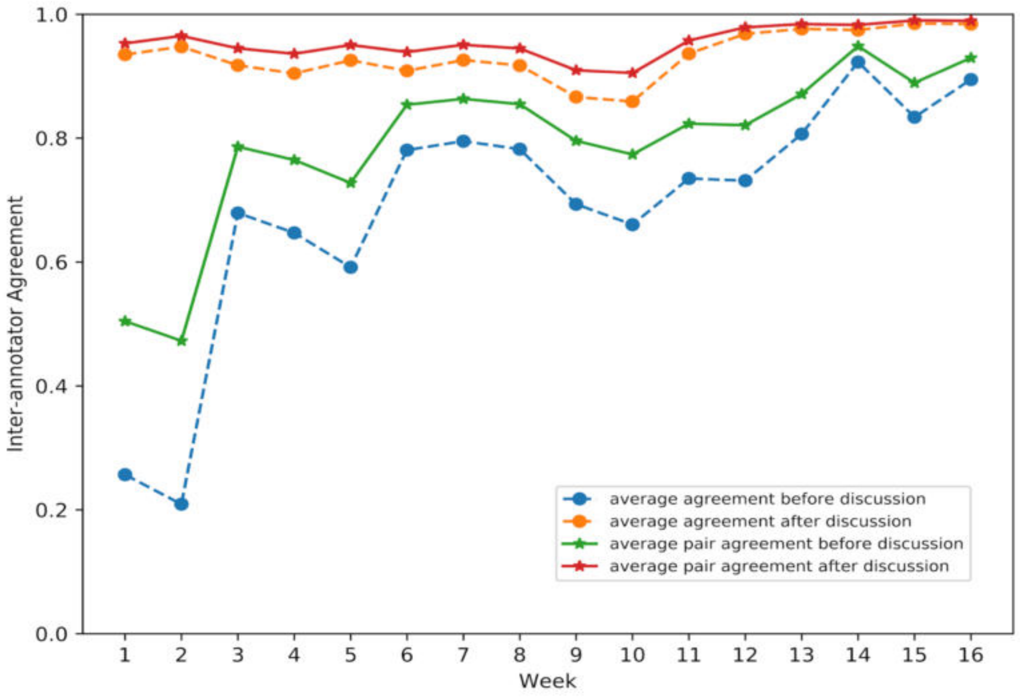
Wang, Mengdi, Hung Chau, Khushboo Thaker, Peter Brusilovsky, and Daqing He. "Knowledge Annotation for Intelligent Textbooks." *Technology, Knowledge and Learning* (2021): 1-22.

Dataset

▷ Section-level concept index for the first 16 chapters of the book Introduction to Information Retrieval (IIR)

Characteristic	
Number of chapters	16
Number of sections	86
Number of all concepts	3175
Number of 1-grams	1121 (35.31%)
Number of 2-grams	1565 (49.29%)
Number of 3-grams	422 (13.29%)
Number of 4-grams	58 (1.83%)
Number of 5+6-grams	9 (0.28%)
Number of all unique concepts	1543
Number of unique 1-grams	278 (18.02%)
Number of unique 2-grams	871 (56.45%)
Number of unique 3-grams	330 (21.39%)
Number of unique 4-grams	55 (3.56%)
Number of unique 5+6-grams	9 (0.58%)

The statistics of the dataset

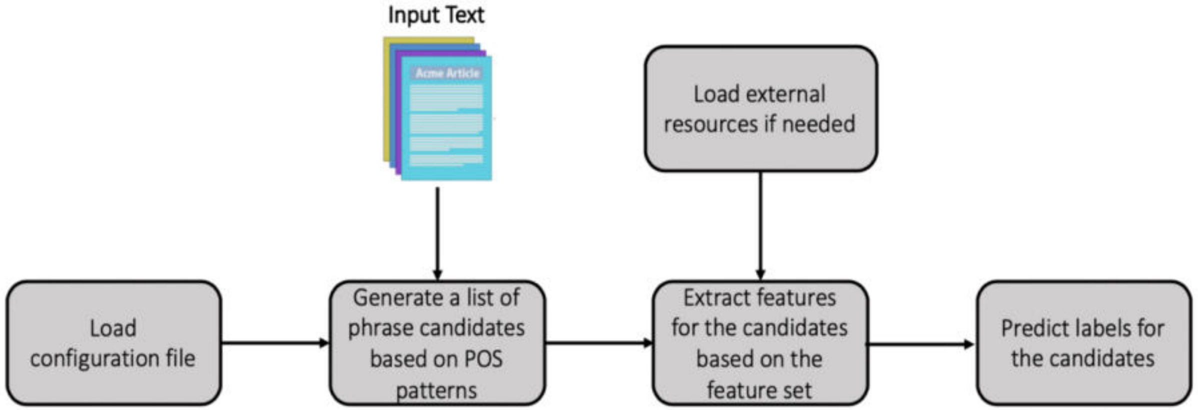


The inter-annotator proportion agreement results (week by week)

FACE: Feature-based Keyphrase Extraction

- Recast as a binary classification problem for a list of extracted candidates
- Candidates extracted based on POS patterns
- Trained a Logistic Regression model with the feature list:
  - Linguistic: POS (features 1-5), context (features 6-17), length of candidate
  - Statistical: frequency, collection frequency, tf-idf, language model
  - External resources: Wikipedia titles, ACM Computer Science keyphrase repository
  - Section titles

Non-binary numerical features are binned and discretized and represented as one-hot encodings



System	AUC	Micro p	Micro r	Micro F_1	Macro p	Macro r	Macro F_1
Baseline 1: Random (*)	0.50	0.20	0.67	0.31	0.14	0.50	0.21
Baseline 2: Linguistics (*)	0.90	0.66	0.63	0.65	0.47	0.57	0.51
Baseline 3: Statistics (*)	0.80	0.55	0.58	0.56	0.50	0.25	0.34
Baseline 4: External resources (*)	0.72	0.36	0.47	0.41	0.29	0.44	0.35
Baseline 5: Titles (*)	0.67	0.55	0.42	0.47	0.43	0.14	0.21
Baseline 6: TextRank	–	0.17	0.10	0.17	0.11	0.10	0.10
Baseline 7: TopicRank	–	0.16	0.16	0.16	0.11	0.28	0.16
Baseline 8: RAKE	–	0.15	0.13	0.14	0.07	**0.63**	0.12
Baseline 10: CopyRNN	–	0.23	0.22	0.23	0.26	0.20	0.23
Baseline 11: Humans (AMT)	–	0.40	0.38	0.39	0.29	0.55	0.38
FACE	**0.94**	**0.75**	**0.77**	**0.76**	**0.61**	0.58	**0.60**

Takeaway messages:

- Book’s rich structure provides useful features to identify keyphrases
- FACE framework can be applied for other domains

Chau, Hung,
concept extra
Journal of Ar

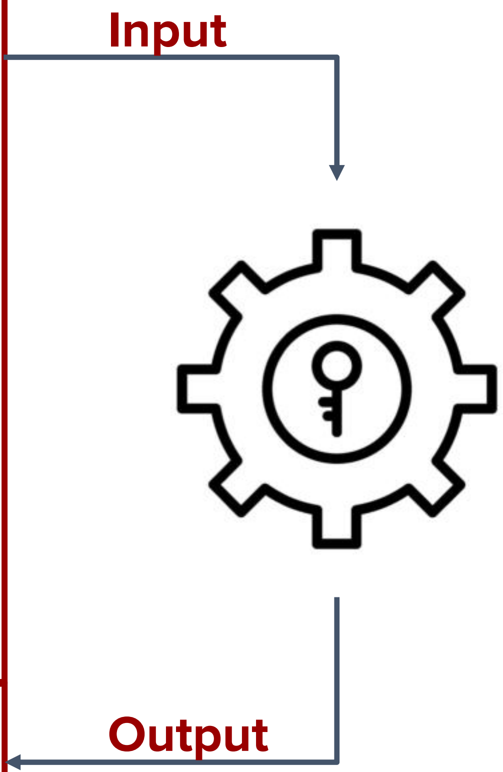
ction

Limitation of Keyphrase Extraction

Muslim Women in Hijab Break Barriers: 'Take the Good With the Bad'

When Ginella Massa, a Toronto-based TV reporter, recently accepted a request to host an evening newscast, she was not planning or expecting to make history for wearing a hijab. She was just covering for a colleague who wanted to go to a hockey game. And that's how Ms. Massa, who works at CityNews in Toronto, became the first Canadian woman to host a newscast from a large media company while wearing the head scarf. [...] This new trend of inclusion occurs amid a more sinister one, as reported hate crimes against Muslims are on the rise in the United States and Canada. The F.B.I. says that a surge in hate crimes against Muslims has led to an overall increase in hate crimes in the United States; Muslims have borne the brunt of the increase with 257 recorded attacks. [...] In Canada, where Ms. Massa has lived since she was a year old, the number of reported hate crimes has dropped slightly overall, but the number of recorded attacks against Muslims has grown: 99 attacks were reported in 2014, according to an analysis by the news site Global News of data from Statistics Canada, a government agency. [...]

keywords: US; Islam; Fashion; Muslim Veiling; Women and Girls; (News media, journalism); Hate crime; Canada



Gallina, Y., Boudin, F., & Daille, B. (2019). KPTimes: A large-scale dataset for keyphrase generation on news documents. arXiv preprint arXiv:1911.12559.

Present Keyphrases (Extractable)

Muslim Women in Hijab Break Barriers: 'Take the Good With the Bad'

When Ginella Massa, a Toronto-based TV reporter, recently accepted a request to host an evening newscast, she was not planning or expecting to make history for wearing a hijab. She was just covering for a colleague who wanted to go to a hockey game. And that's how Ms. Massa, who works at CityNews in Toronto, became the first Canadian woman to host a newscast from a large media company while wearing the head scarf. [...] This new trend of inclusion occurs amid a more sinister one, as reported hate crimes against Muslims are on the rise in the United States and Canada. The F.B.I. says that a surge in hate crimes against Muslims has led to an overall increase in hate crimes in the United States; Muslims have borne the brunt of the increase with 257 recorded attacks. [...] In Canada, where Ms. Massa has lived since she was a year old, the number of reported hate crimes has dropped slightly overall, but the number of recorded attacks against Muslims has grown: 99 attacks were reported in 2014, according to an analysis by the news site Global News of data from Statistics Canada, a government agency. [...]

keywords: US; Islam; Fashion; Muslim Veiling; Women and Girls; (News media, journalism); Hate crime; Canada

Input

Output

Keyphrases (or part of) appearing in the document are colored

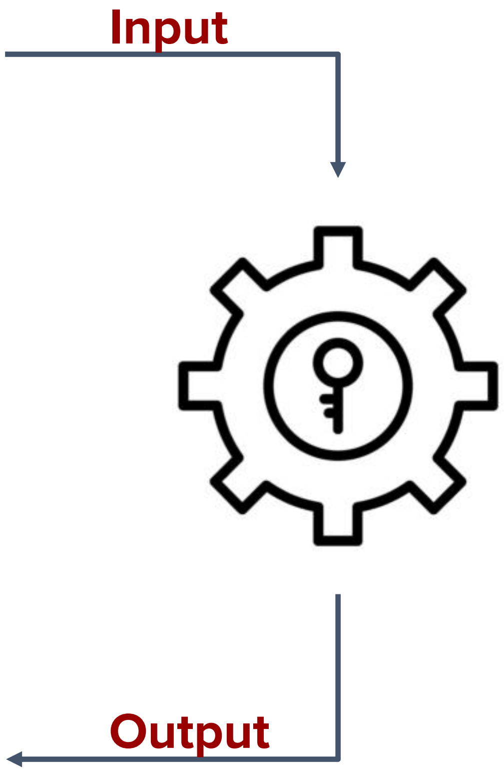
Gallina, Y., Boudin, F., & Daille, B. (2019). KPTimes: A large-scale dataset for keyphrase generation on news documents. arXiv preprint arXiv:1911.12559.

Absent Keyphrases (Not Extractable)

Muslim Women in Hijab Break Barriers: 'Take the Good With the Bad'

When Ginella Massa, a Toronto-based TV reporter, recently accepted a request to host an evening newscast, she was not planning or expecting to make history for wearing a hijab. She was just covering for a colleague who wanted to go to a hockey game. And that's how Ms. Massa, who works at CityNews in Toronto, became the first Canadian woman to host a newscast from a large media company while wearing the head scarf. [...] This new trend of inclusion occurs amid a more sinister one, as reported hate crimes against Muslims are on the rise in the United States and Canada. The F.B.I. says that a surge in hate crimes against Muslims has led to an overall increase in hate crimes in the United States; Muslims have borne the brunt of the increase with 257 recorded attacks. [...] In Canada, where Ms. Massa has lived since she was a year old, the number of reported hate crimes has dropped slightly overall, but the number of recorded attacks against Muslims has grown: 99 attacks were reported in 2014, according to an analysis by the news site Global News of data from Statistics Canada, a government agency. [...]

keywords: US; Islam; Fashion; Muslim Veiling; Women and Girls; (News media, journalism); Hate crime; Canada



Keyphrases (or part of) appearing in the document are colored

Gallina, Y., Boudin, F., & Daille, B. (2019). KPTimes: A large-scale dataset for keyphrase generation on news documents. arXiv preprint arXiv:1911.12559.

Not All Keyphrases Are Extractable

- A non-negligible proportion of keyphrases are not present
  - Annotators assign keyphrases by their relevance/importance, not presence

Dataset	#Train	#Valid	#Test	Mean	Var	%Pre
KP20K	≈514k	≈20k	≈20k	5.3	14.2	63.3%
INSPEC	–	1500	500	9.6	22.4	78.5%
KRAPIVIN	–	1844	460	5.2	6.6	56.2%
NUS	–	-	211	11.5	64.6	51.3%
SEMEVAL	–	144	100	15.7	15.1	44.5%
STACKEX	≈298k	≈16k	≈16k	2.7	1.4	57.5%

(Yuan et al. 2018). One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases. ACL.

Method 2: Neural Keyphrase Generation

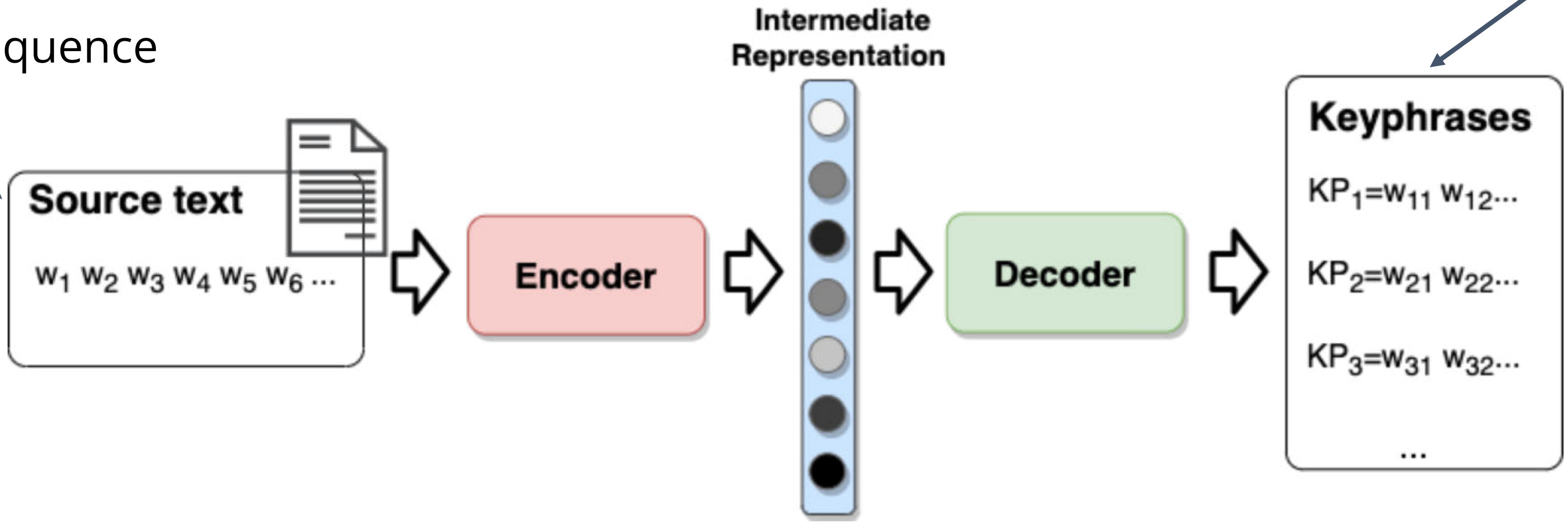
▷ Predicting keyphrases as language generation

- Each keyphrase is actually a short sequence of tokens
- We can train neural networks to learn to generate phrases in a data-driven way

Input: a **SEQ**uence of source text
Output: multiple **SEQ**uences of tokens, each sequence is a keyphrase
Seq2Seq Learning!

Multiple target sequences

One source sequence

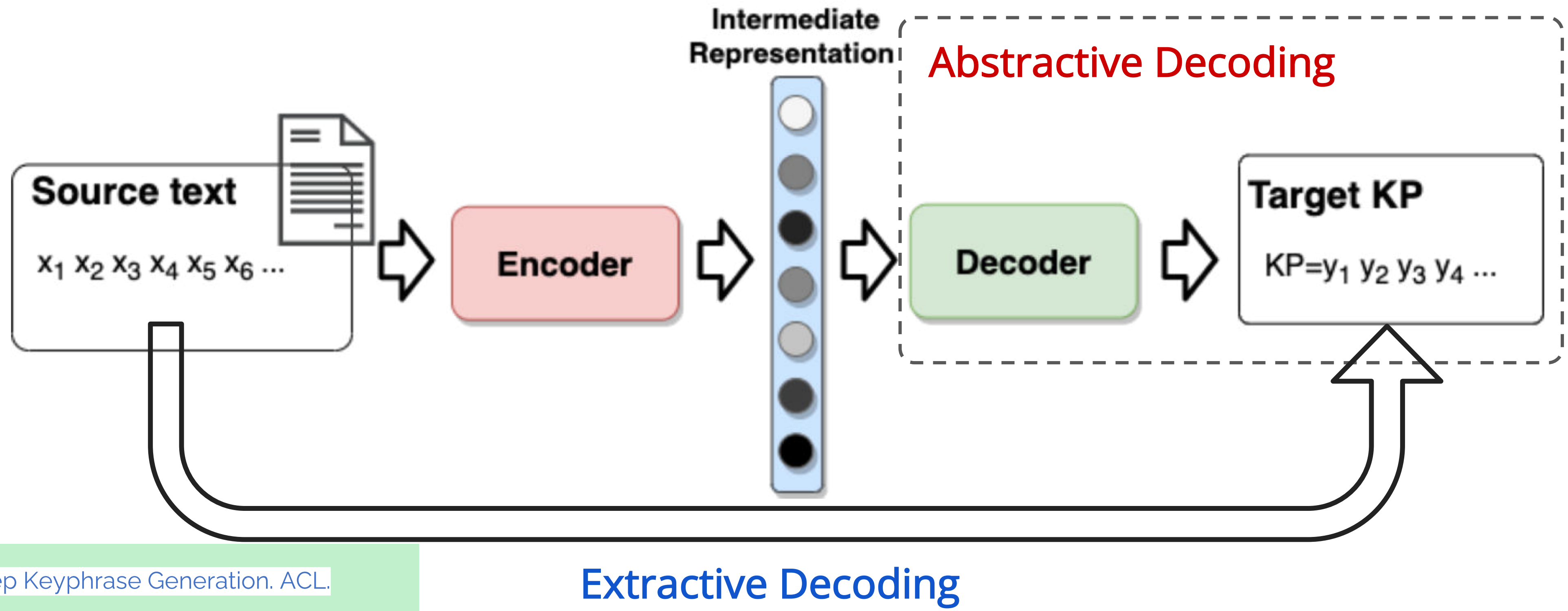


Keyphrase Generation Models

- Seq2Seq + Copy Attention

  Generate target keyphrase both abstractively and extractively

$$P(w)= p_{abs} * P_{abs}(w_{vocab}) + (1 - p_{abs}) * P_{ext}(w_{src})$$



(Meng et al. 2017). Deep Keyphrase Generation. ACL.

Keyphrase Generation (KPG)

- Three types of training paradigms

  One2One: Output <u>one single phrase</u> at a time

  One2Seq: Output <u>a sequence of multiple phrases</u> at a time

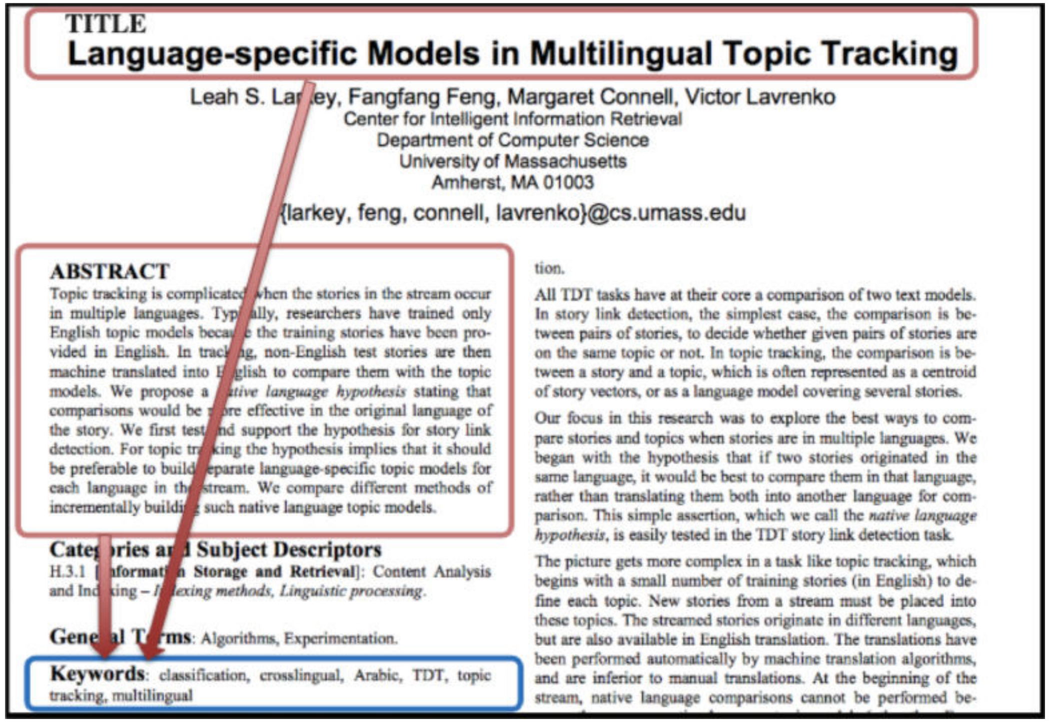
  One2Set: Output <u>a set of multiple phrases</u> at a time

Rui Meng, Debanjan Mahata, Florian Boudin “From Fundamentals to Recent Advances: A Tutorial on Keyphrasification”, a half-day tutorial at the 44th European Conference on Information Retrieval (ECIR 2022) https://keyphrasification.github.io/

(Meng et al. 2017). Deep Keyphrase Generation. ACL.
(Yuan et al. 2018). One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases. ACL.
(Ye and Wang, 2018). Semi-Supervised Learning for Neural Keyphrase Generation. EMNLP.
(Meng et al. 2021). An Empirical Study on Neural Keyphrase Generation. NAACL.
(Ye et al. 2021) "One2Set: Generating Diverse Keyphrases as a Set. ACL.

KPG-One2One Vs. KPG-One2Seq

- Both are based on Sequence-to-Sequence Learning



TITLE

Language-specific Models in Multilingual Topic Tracking

Leah S. Lar ey, Fangfang Feng, Margaret Connell, Victor Lavrenko
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{larkey, feng, connell, lavrenko}@cs.umass.edu

ABSTRACT

Topic tracking is complicated hen the stories in the stream occur in multiple languages. Typ ally, researchers have trained only English topic models beca e the training stories have been provided in English. In track ng, non-English test stories are then machine translated into E glish to compare them with the topic models. We propose a ative language hypothesis stating that comparisons would be ore effective in the original language of the story. We first test nd support the hypothesis for story link detection. For topic tr king the hypothesis implies that it should be preferable to build eparate language-specific topic models for each language in th stream. We compare different methods of incrementally buildi such native language topic models.

Categ ries a d Subject Descriptors
H.3.1 [nformat n **Storage and Retrieval**]: Content Analysis and Ind xing – *I exing methods, Linguistic processing*.

Gene al T rms: Algorithms, Experimentation.

Keywords: classification, crosslingual, Arabic, TDT, topic tracking, multilingual

tion.

All TDT tasks have at their core a comparison of two text models. In story link detection, the simplest case, the comparison is between pairs of stories, to decide whether given pairs of stories are on the same topic or not. In topic tracking, the comparison is between a story and a topic, which is often represented as a centroid of story vectors, or as a language model covering several stories.

Our focus in this research was to explore the best ways to compare stories and topics when stories are in multiple languages. We began with the hypothesis that if two stories originated in the same language, it would be best to compare them in that language, rather than translating them both into another language for comparison. This simple assertion, which we call the *native language hypothesis*, is easily tested in the TDT story link detection task.

The picture gets more complex in a task like topic tracking, which begins with a small number of training stories (in English) to define each topic. New stories from a stream must be placed into these topics. The streamed stories originate in different languages, but are also available in English translation. The translations have been performed automatically by machine translation algorithms, and are inferior to manual translations. At the beginning of the stream, native language comparisons cannot be performed be-

[Source Sequence]=title+abstract

Language-specific Models in Multilingual Topic Tracking. Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare them with the topic models. …

[Target Sequence]=a list of keyphrases

[classification, crosslingual, Arabic, TDT, topic tracking, multilingual]

KPG-One2One

- Data preparation - each data example is split to multiple text-keyphrase pairs
  - Source text is duplicated K times
  - Each pair contains only one keyphrase
  - Great waste in training, e.g. in KP20k 510K->2.78M

Original Data Point (k target phrases)

[Source]
Language-specific Models in Multilingual Topic Tracking. Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare them with the topic models. …

[Target]
[**classification**, crosslingual, Arabic, TDT, topic tracking, multilingual]

Src-Tgt Pair for Training (k pairs)

[Source] Language-specific Models in Multilingual Topic Tracking.…
[Target] <s> classification </s>

[Source] Language-specific Models in Multilingual Topic Tracking.…
[Target] <s> crosslingual </s>

[Source] Language-specific Models in Multilingual Topic Tracking.…
[Target] <s> arabic </s>

[Source] Language-specific Models in Multilingual Topic Tracking.…
[Target] <s> TDT </s>

[Source] Language-specific Models in Multilingual Topic Tracking.…
[Target] <s> topic tracking </s>

[Source] Language-specific Models in Multilingual Topic Tracking.…
[Target] <s> multilingual </s>

(Meng et al. 2017). Deep Keyphrase Generation. ACL.

KPG-One2Seq

- Can the model generate multiple phrases directly?
  - Output
  - Let model to handle the interaction between phrases and avoid redundancy in output
- KPGen-One2Seq
  - Given **ONE** source text, the goal is to generate one **SEQ**uence of concatenated keyphrases
  - Concatenate multiple target phrases as a sequence
  - The order of concatenation can be effective in performance

[Source Sequence]
Language-specific Models in Multilingual Topic Tracking. Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare them with the topic models. …

[Target Sequence]
[**classification**, crosslingual, Arabic, TDT, topic tracking, multilingual]

→

[Source] Language-specific Models in Multilingual Topic Tracking.…
[Target] <s> classification <sep> crosslingual <sep> Arabic <sep> TDT <sep> topic tracking <sep> multilingual </s>

(Yuan et al. 2018). One Size Does Not Fit All: Generating and Evaluating Variable Number of Keyphrases. ACL.
(Ye and Wang, 2018). Semi-Supervised Learning for Neural Keyphrase Generation. EMNLP.

KPG-One2Seq

- One2Seq: But the order of concatenation may matter ...
- Several order options

[Source Sequence]=title+abstract
Language-specific Models in Multilingual Topic Tracking. Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare them with the topic models. ...

[Target Sequence]=keyphrases
[classification, crosslingual, Arabic, TDT, topic tracking, multilingual]

[Present Phrases] topic tracking, multilingual
[Absent Phrases] classification, crosslingual, Arabic, TDT

Random	[Source] Language-specific Models in Multilingual Topic [Target] <bos> TDT <sep> multilingual <sep> crosslingual <sep> Arabic <sep> classification <sep> topic tracking
Length	[Source] Language-specific Models in Multilingual Topic [Target] <bos> classification <sep> crosslingual <sep> Arabic <sep> TDT <sep> multilingual <sep> topic tracking
No-Sort	[Source] Language-specific Models in Multilingual Topic [Target] <bos> classification <sep> crosslingual <sep> Arabic <sep> TDT <sep> topic tracking <sep> multilingual
Alpha	[Source] Language-specific Models in Multilingual Topic [Target] <bos> Arabic <sep>classification <sep> crosslingual <sep> multilingual <sep> TDT <sep> topic tracking
Absent-PRE	[Source] Language-specific Models in Multilingual Topic [Target] <bos> Arabic <sep> TDT <sep>classification <sep> crosslingual <sep> multilingual <sep> topic tracking
Absent-AP	[Source] Language-specific Models in Multilingual Topic [Target] <bos> multilingual <sep> topic tracking <sep> TDT <sep> Arabic <sep> classification <sep> crosslingual

Datasets

- Datasets
  - KP20k (514k CS papers)
  - Inspec: 2,000 paper abstracts.
  - Krapivin: 2,304 papers with full-text and author-assigned keyphrases.
  - NUS: 211 papers with author-assigned and reader-assigned Keyphrases
  - SemEval-2010: 288 articles from the ACM Digital Library
- Relations
  Same domain, similar distribution
  - KP20k, Krapivin

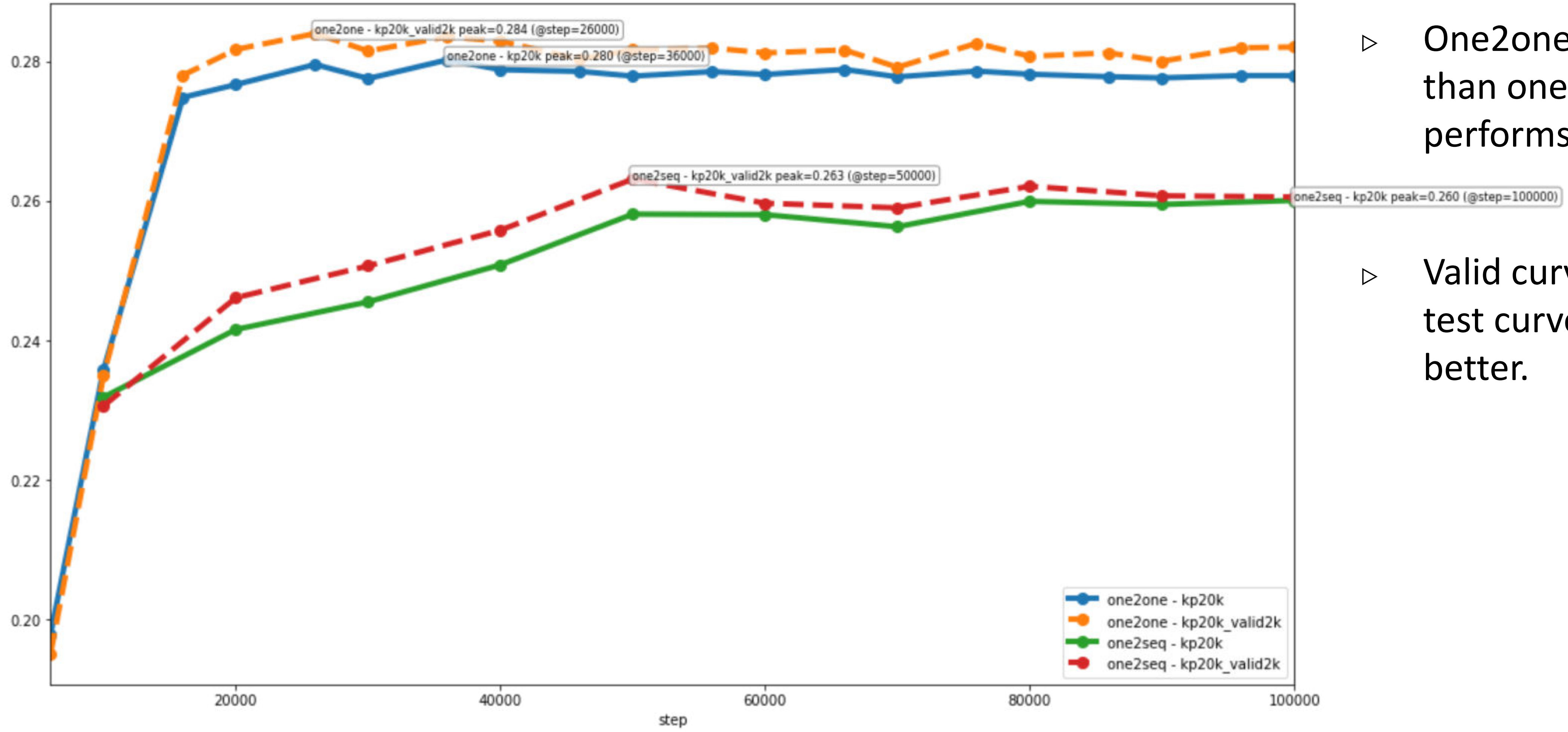
  Same domain, different distribution/annotation
  - Inspec, NUS, SemEval

  Different domain
  - DUC (news article)

Dataset	#Train	#Valid	#Test	Mean	Var	%Pre
KP20K	≈514k	≈20k	≈20k	5.3	14.2	63.3%
MAGKP	≈2.7M	-	-	12.9	?	?%
INSPEC	–	1500	500	9.6	22.4	78.5%
KRAPIVIN	–	1844	460	5.2	6.6	56.2%
NUS	–	-	211	11.5	64.6	51.3%
SEMEVAL	–	144	100	15.7	15.1	44.5%
STACKEX	≈298k	≈16k	≈16k	2.7	1.4	57.5%
DUC	-	-	308	8.1	?	97.5%

Table 1: Statistics of various datasets. Mean and Var indicate the mean and variance of target phrase numbers, %Pre denotes percentage of present keyphrases.

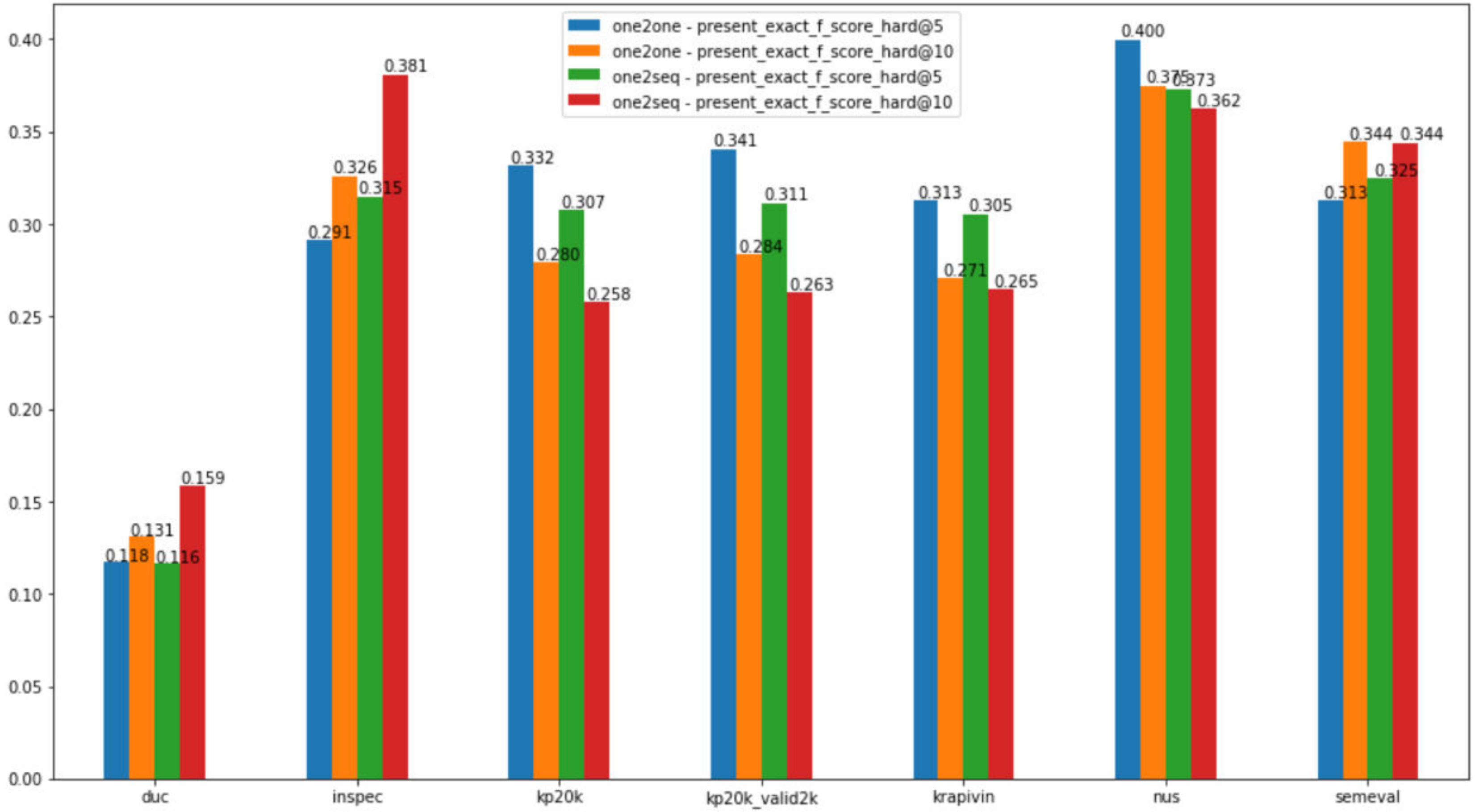
Learning Curve (F@10) of In-Domain Datasets



Impressions

- One2one converges faster than one2seq, and performs better
- Valid curve is in line with test curve, and always better.

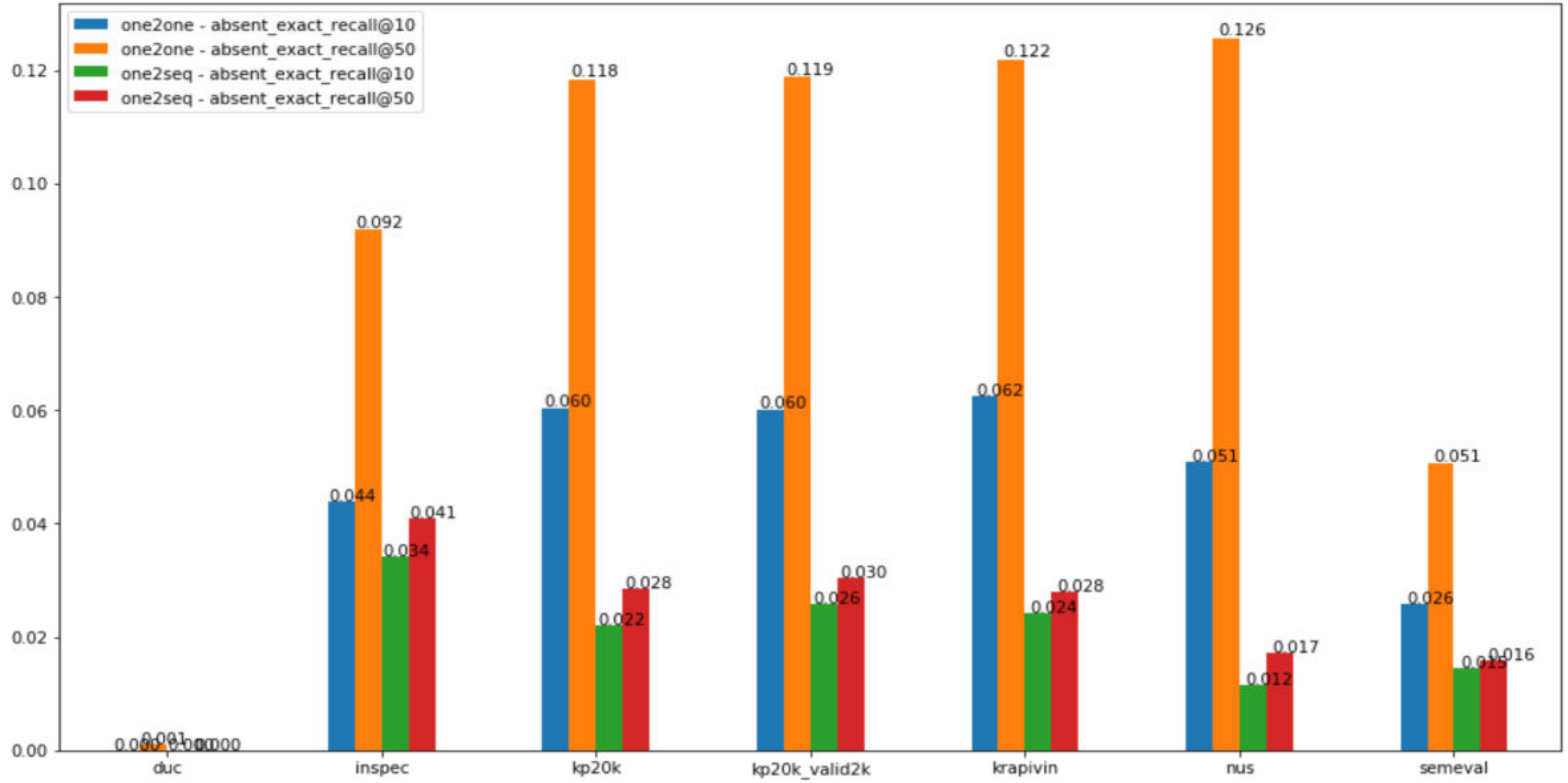
A Closer Look at All Datasets- Present



Impressions

- One2One performs better on in-domain datasets (KP20K and Krapivin) and NUS.
- One2Seq's transferability looks better: on Inspec and DUC (news) it outperforms One2One significantly

A Closer Look at All Datasets - Absent



Impressions

- One2One performs much better than One2Seq due to its superior ability in generating unique phrases.

Takeaway messages:

- Keyphrase generation is a more powerful task modeling
- But its effective generation methods are still opening questions

Outline

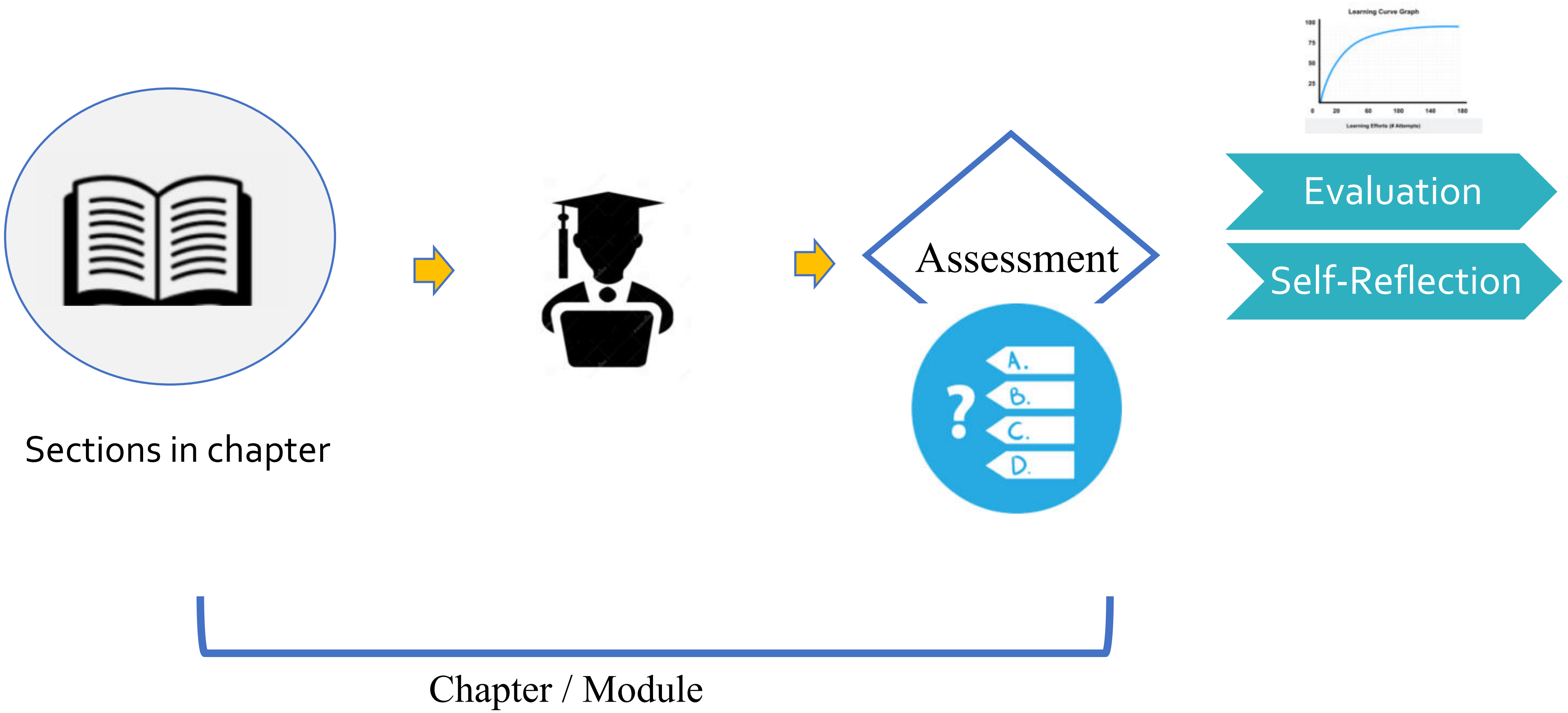
- Basics of Keyphrases: Definitions and Importance
- Identification of Keyphrases: Extraction and Generation
- Applications of Keyphrases: knowledge unit for supporting student learning
- Applications of Keyphrases: knowledge unit for recognizing patients' concerns
- Applications of Keyphrases: knowledge unit for interactive machine learning
- Conclusions

Recommending Remedial Readings Using Student Knowledge State

Khushboo Thaker, Lei Zhang, Daqing He, Peter Brusilovsky

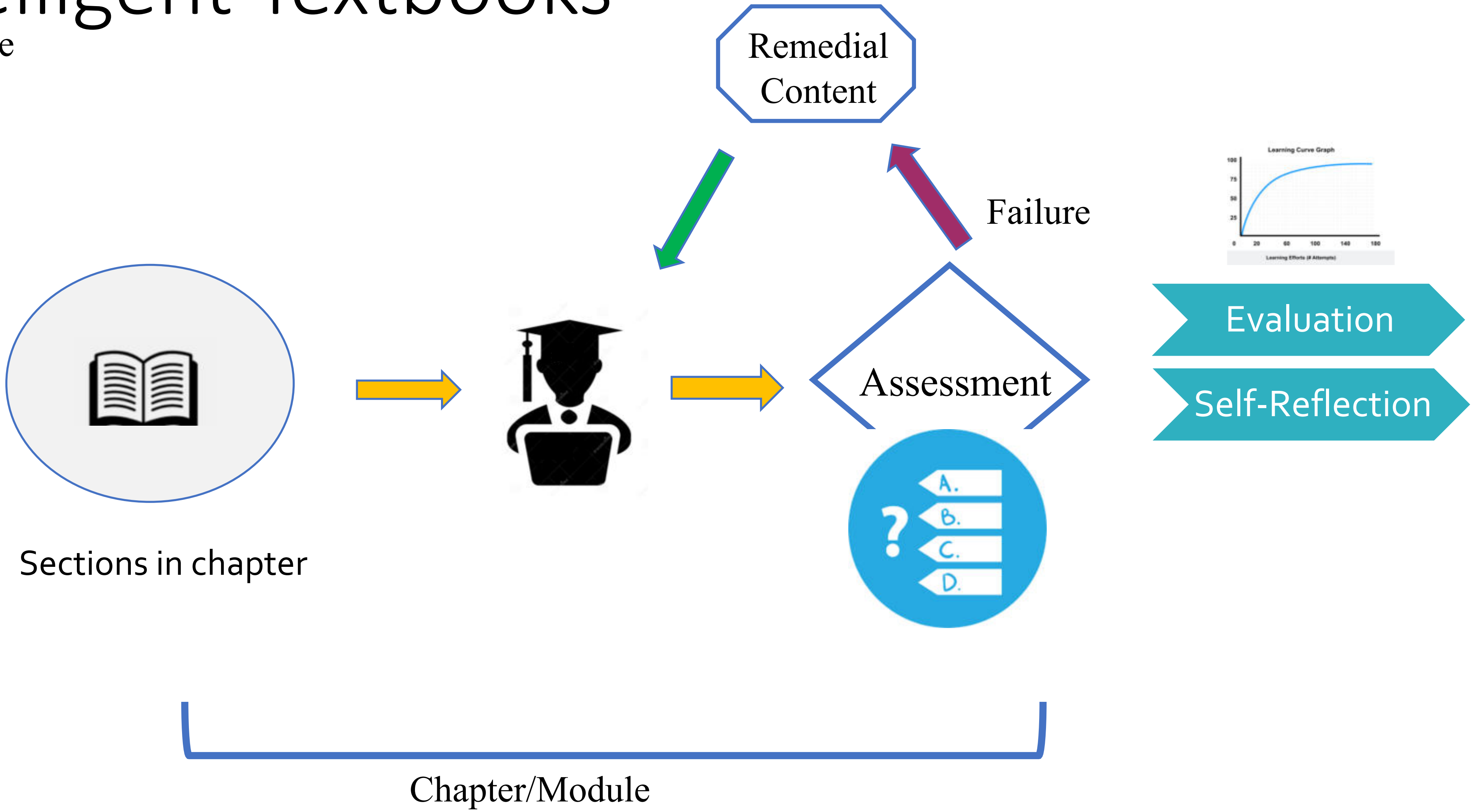
Intelligent Textbooks

Background



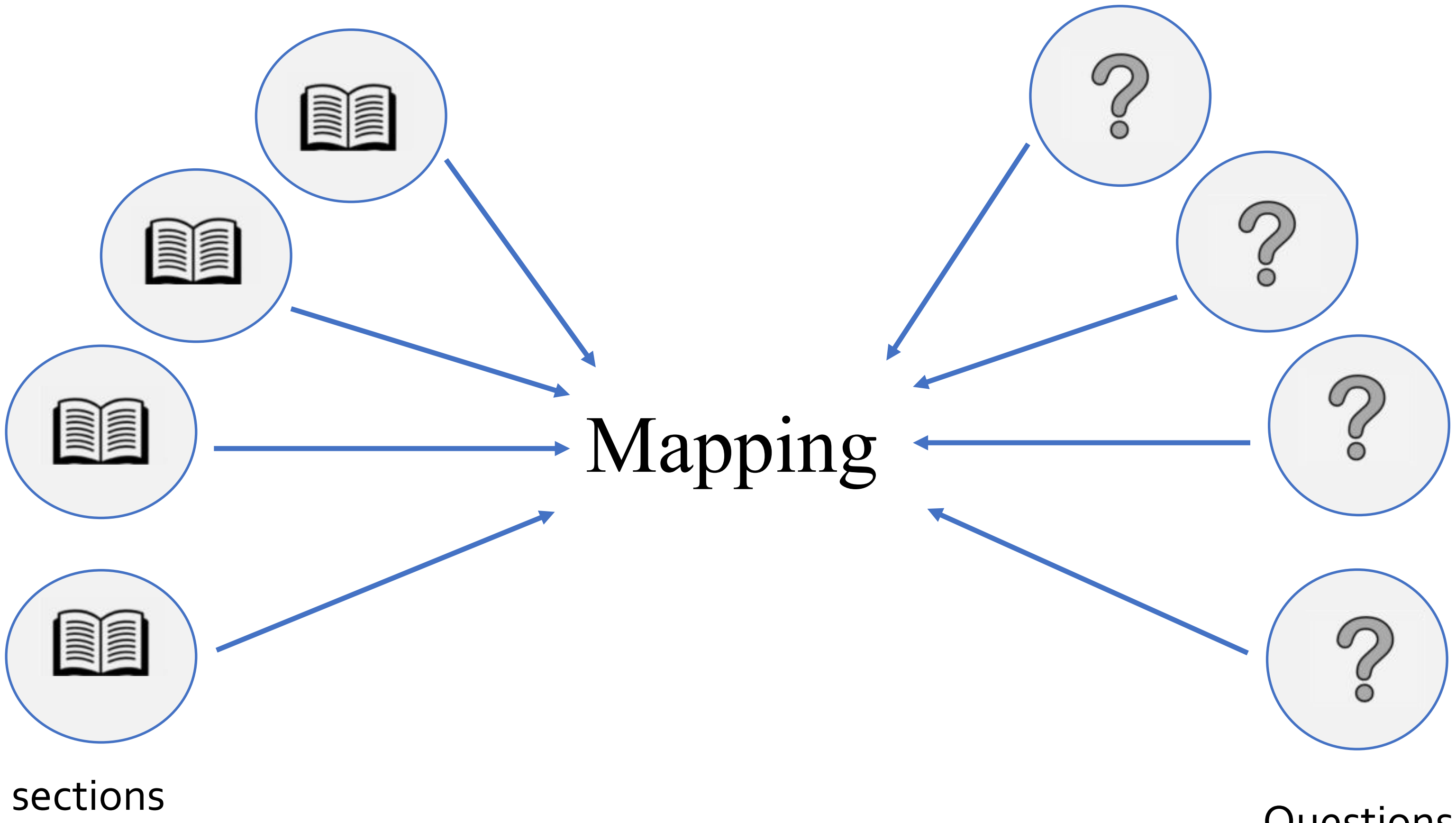
Intelligent Textbooks

objective

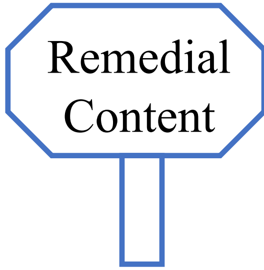


Remedial Recommendation

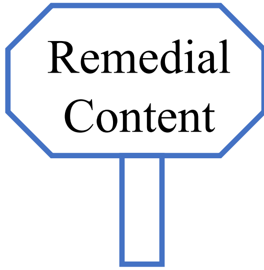
Challenge

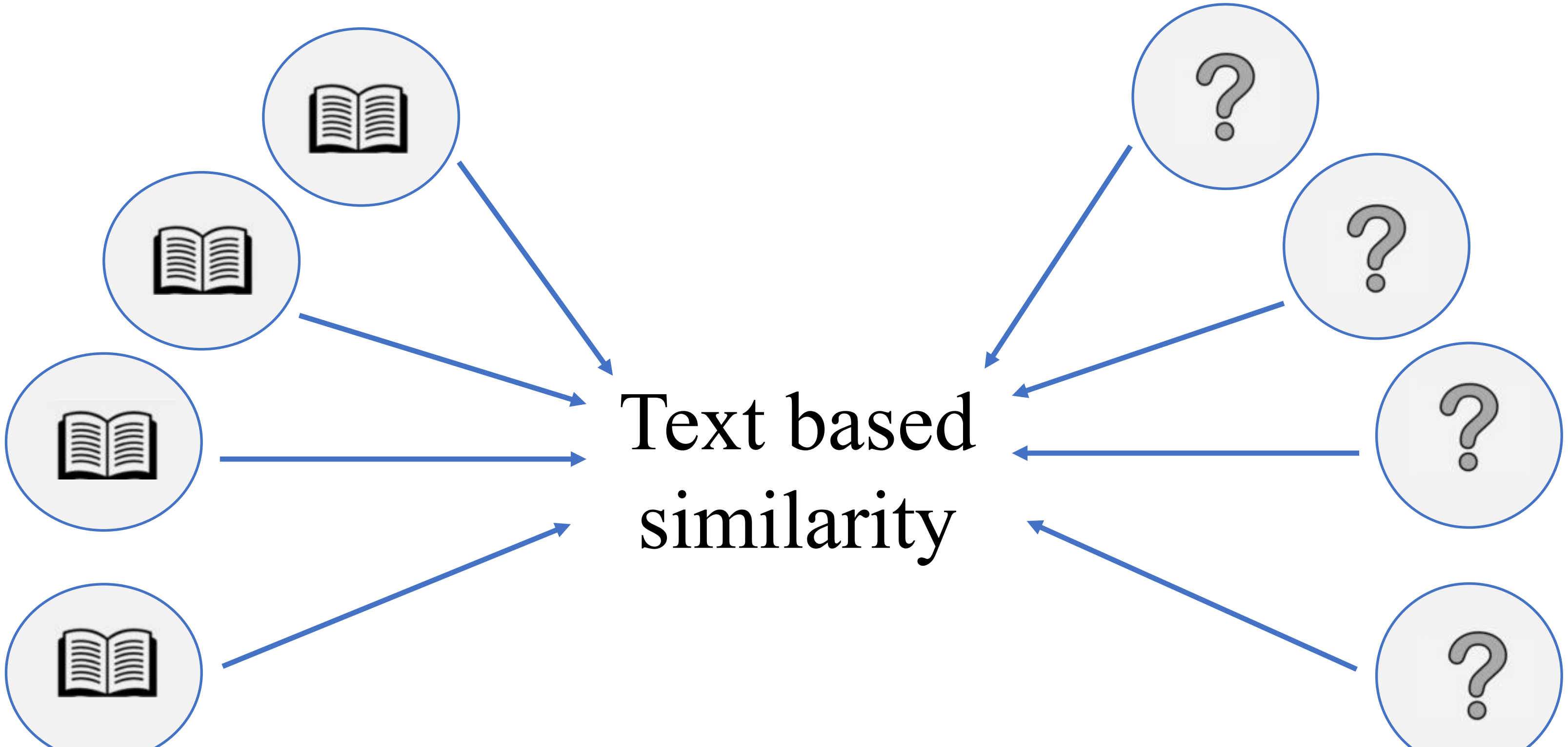


Computer Assisted Instructions (CAI)



Simple Text-based Similarity and its challenge



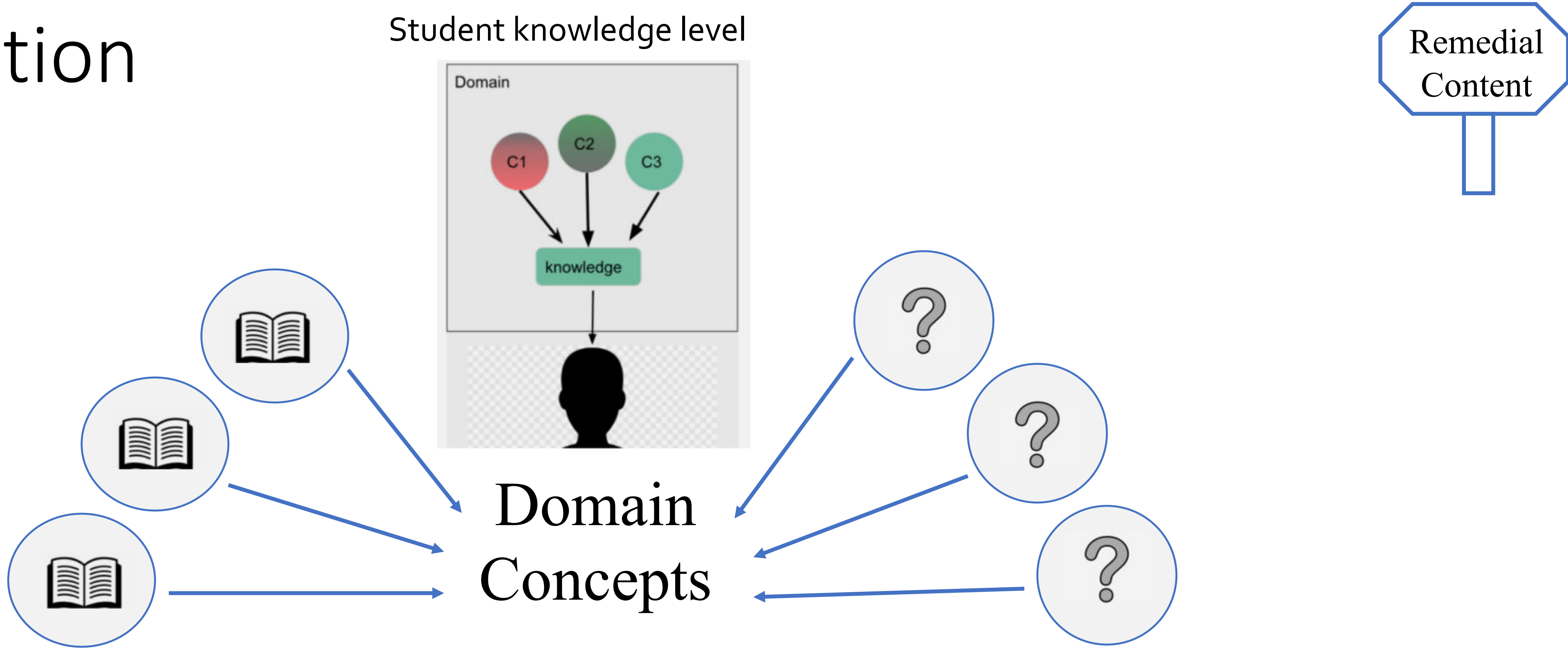


<u>Challenges:</u>

- **Advance Content** - Student lack pre-requisites
- **Redundant Content** - Student already mastered
- **Not Personalized** to students need

B. Pursel, C. Liang, S. Wang, Z. Wu, K. Williams,B. Br¨autigam, S. Saul, H. Williams, K. Bowen, and C. L. Giles. Bbookx: Design of an automated web-based recommender system for the creation of open learning content. WWW 2016

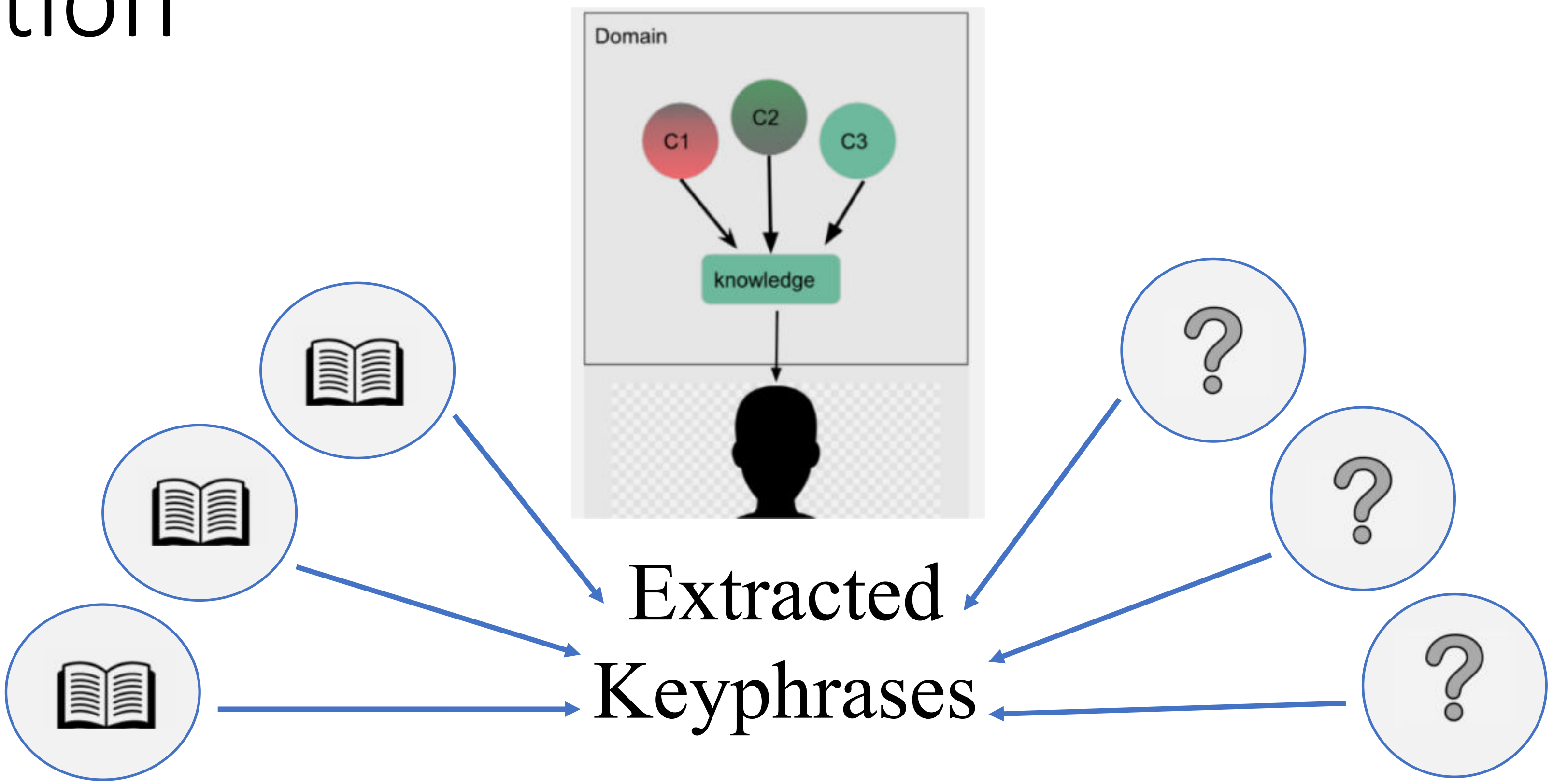
Motivation

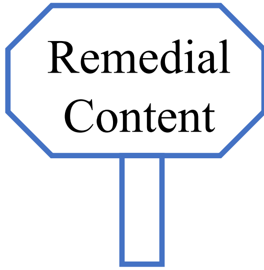


Example Text:
Information retrieval is the activity of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds.

Motivation

Student knowledge level





Example Text:
Information retrieval is the activity of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds.

Research Questions:

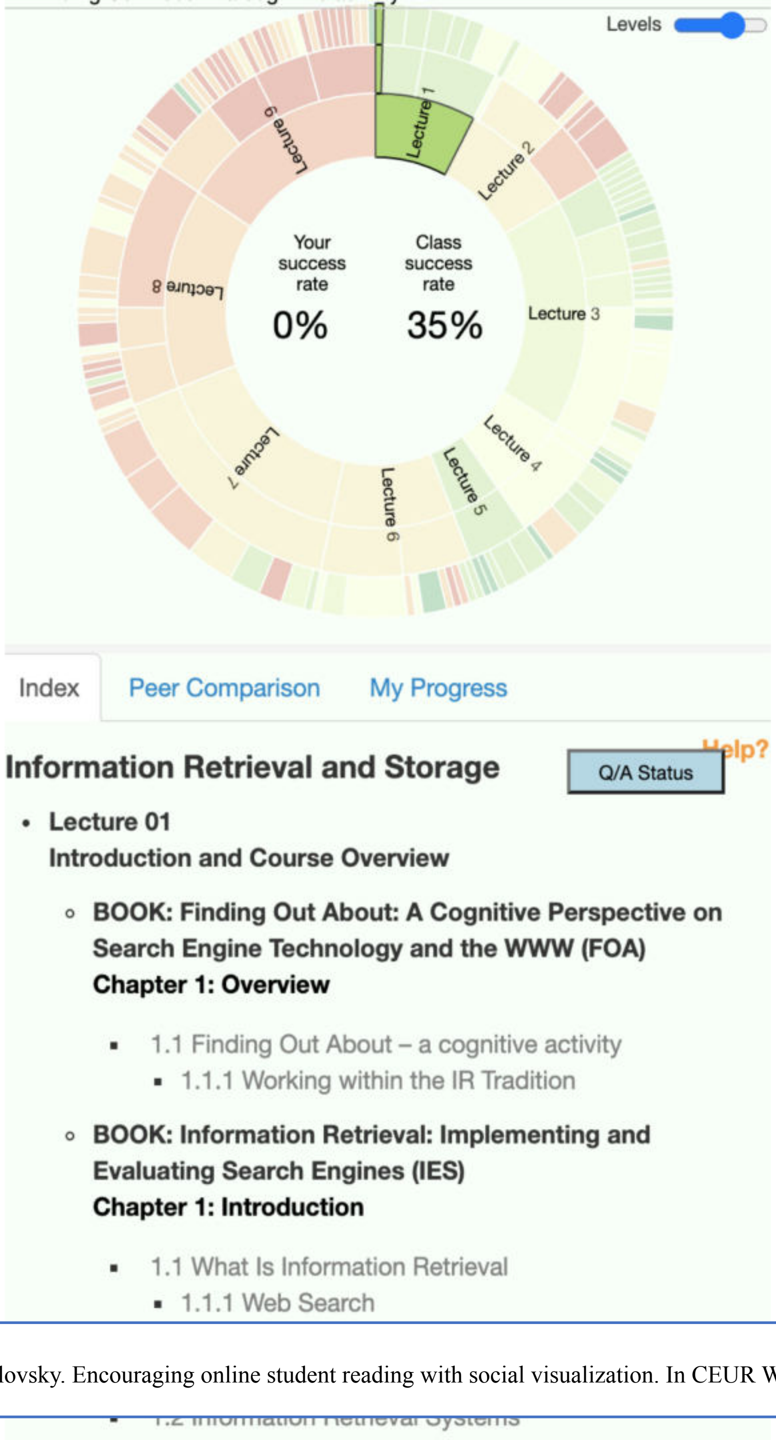
1. Does the **concept-based representation** of educational content help perform remedial recommendation, either by acting alone or in combination with the content-based recommendation?

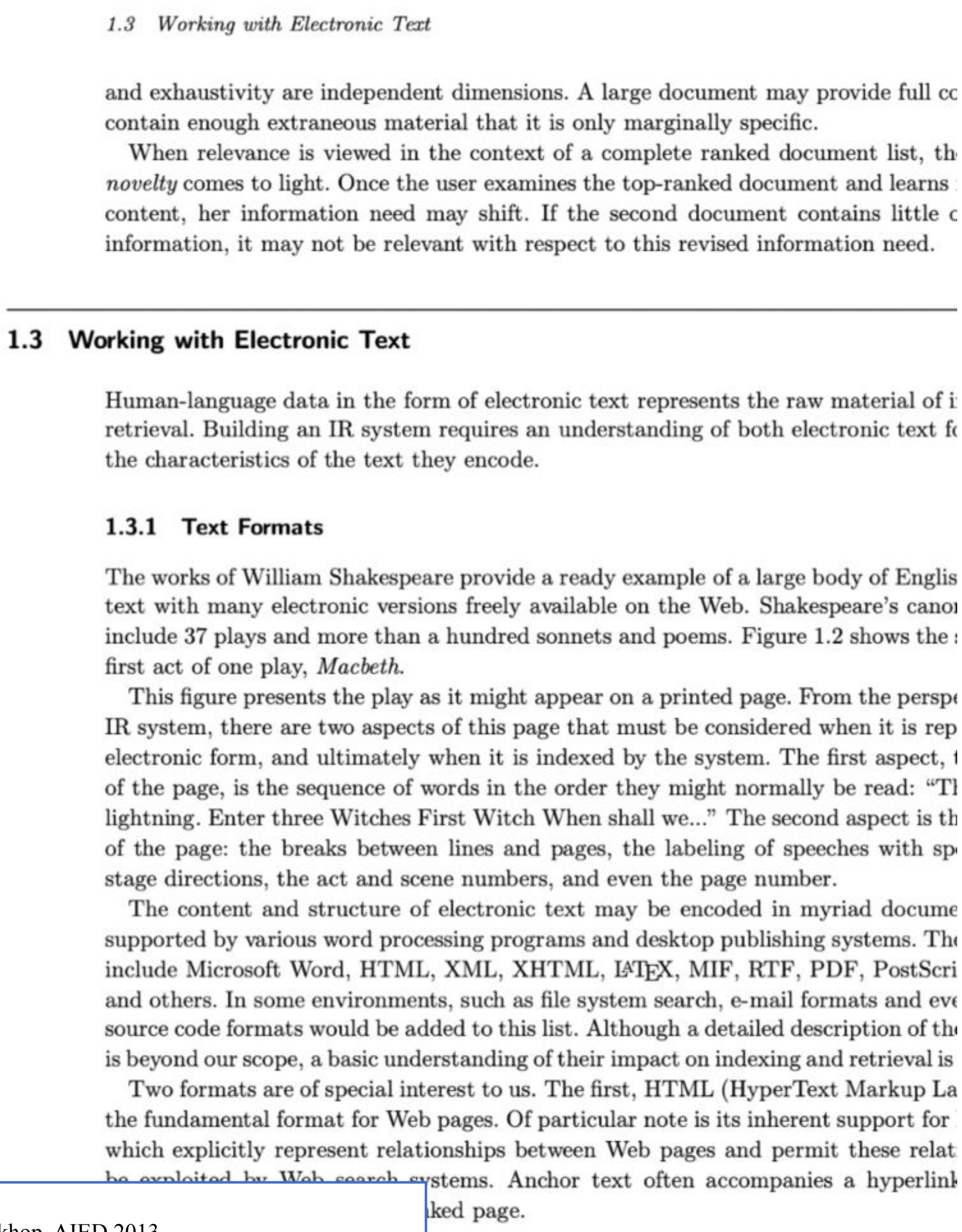
2. Does the augmentation of **student knowledge** on concept-based representation help in provide personalized remedial recommendations?

3 Can we use **automated key phrase extraction** techniques to generate concept-based representation ?

ReadingCircle: Information Retrieval Textbook

Each Section is annotated with identified keyphrases as domain concepts





1.3 Working with Electronic Text

and exhaustivity are independent dimensions. A large document may provide full cc contain enough extraneous material that it is only marginally specific.

When relevance is viewed in the context of a complete ranked document list, th *novelty* comes to light. Once the user examines the top-ranked document and learns content, her information need may shift. If the second document contains little c information, it may not be relevant with respect to this revised information need.

1.3 Working with Electronic Text

Human-language data in the form of electronic text represents the raw material of i retrieval. Building an IR system requires an understanding of both electronic text fc the characteristics of the text they encode.

1.3.1 Text Formats

The works of William Shakespeare provide a ready example of a large body of Englis text with many electronic versions freely available on the Web. Shakespeare's canoı include 37 plays and more than a hundred sonnets and poems. Figure 1.2 shows the first act of one play, *Macbeth*.

This figure presents the play as it might appear on a printed page. From the perspe IR system, there are two aspects of this page that must be considered when it is rep electronic form, and ultimately when it is indexed by the system. The first aspect, of the page, is the sequence of words in the order they might normally be read: "Tl lightning. Enter three Witches First Witch When shall we..." The second aspect is th of the page: the breaks between lines and pages, the labeling of speeches with sp stage directions, the act and scene numbers, and even the page number.

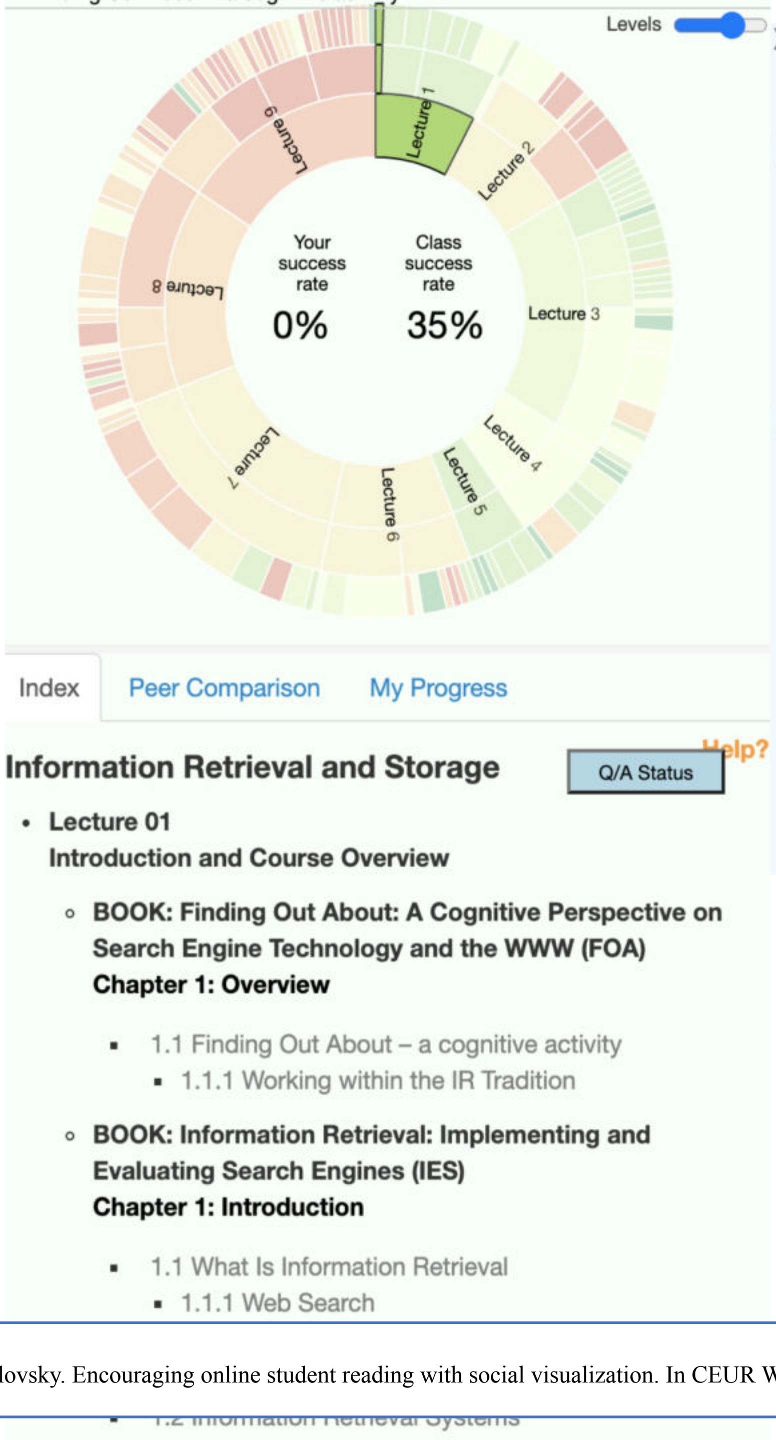
The content and structure of electronic text may be encoded in myriad docume supported by various word processing programs and desktop publishing systems. Th include Microsoft Word, HTML, XML, XHTML, LaTeX, MIF, RTF, PDF, PostScri and others. In some environments, such as file system search, e-mail formats and eve source code formats would be added to this list. Although a detailed description of th is beyond our scope, a basic understanding of their impact on indexing and retrieval is

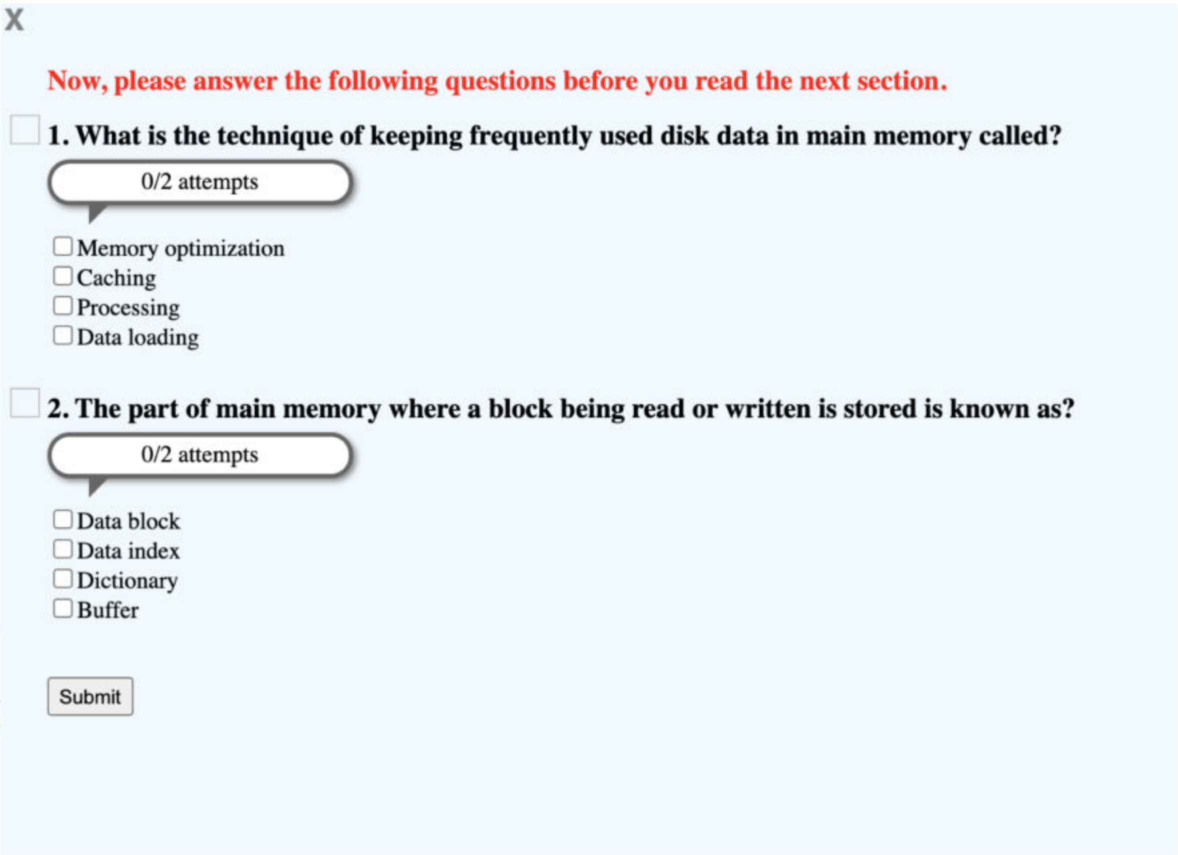
Two formats are of special interest to us. The first, HTML (HyperText Markup La the fundamental format for Web pages. Of particular note is its inherent support for which explicitly represent relationships between Web pages and permit these relat be exploited by Web search systems. Anchor text often accompanies a hyperlink ...ked page.

ReadingCircle: J. Guerra, D. Parra, and P. Brusilovsky. Encouraging online student reading with social visualization. In CEUR Workhop, AIED 2013

ReadingCircle: Information Retrieval Textbook

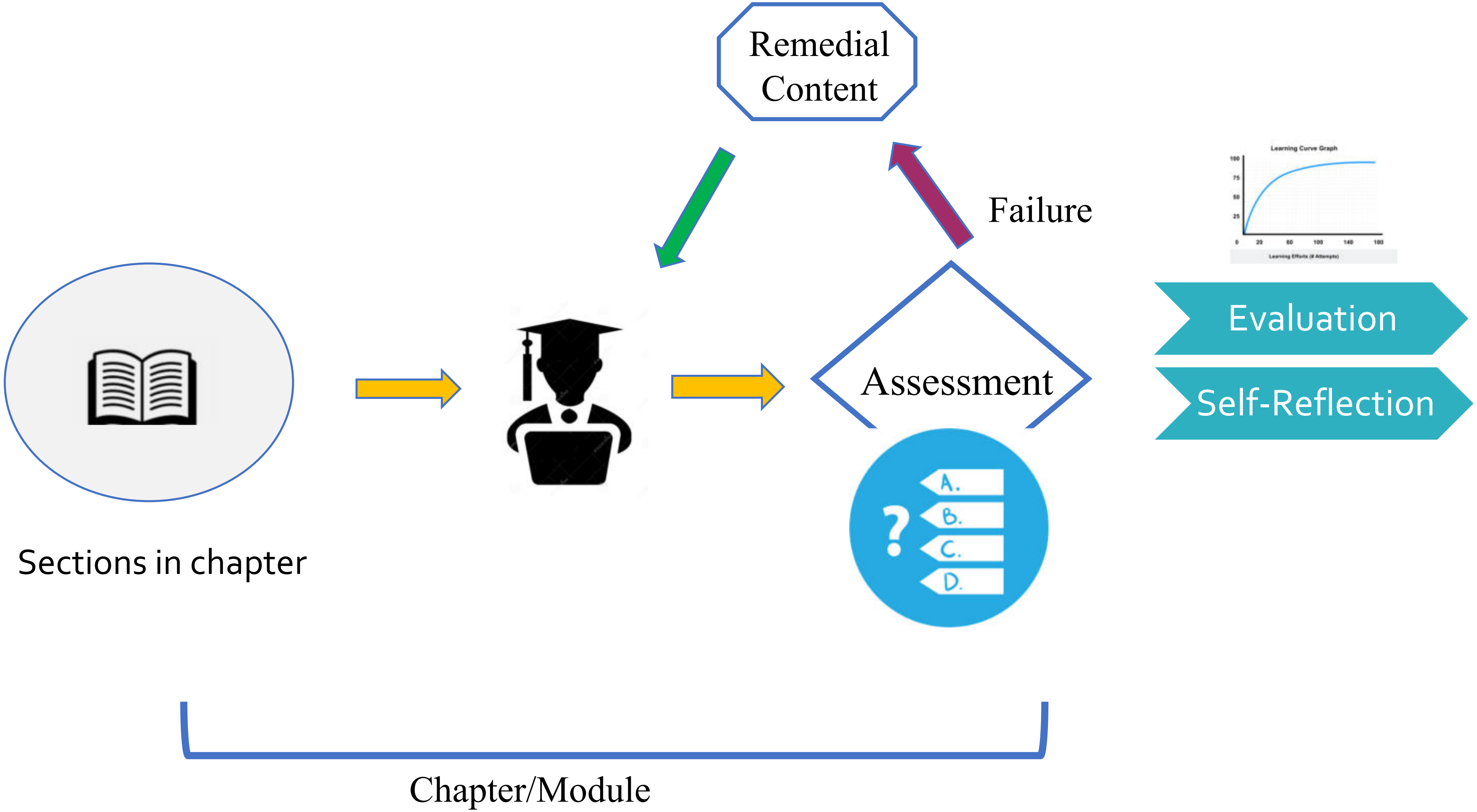
Each Section is annotated with identified keyphrases as domain concepts



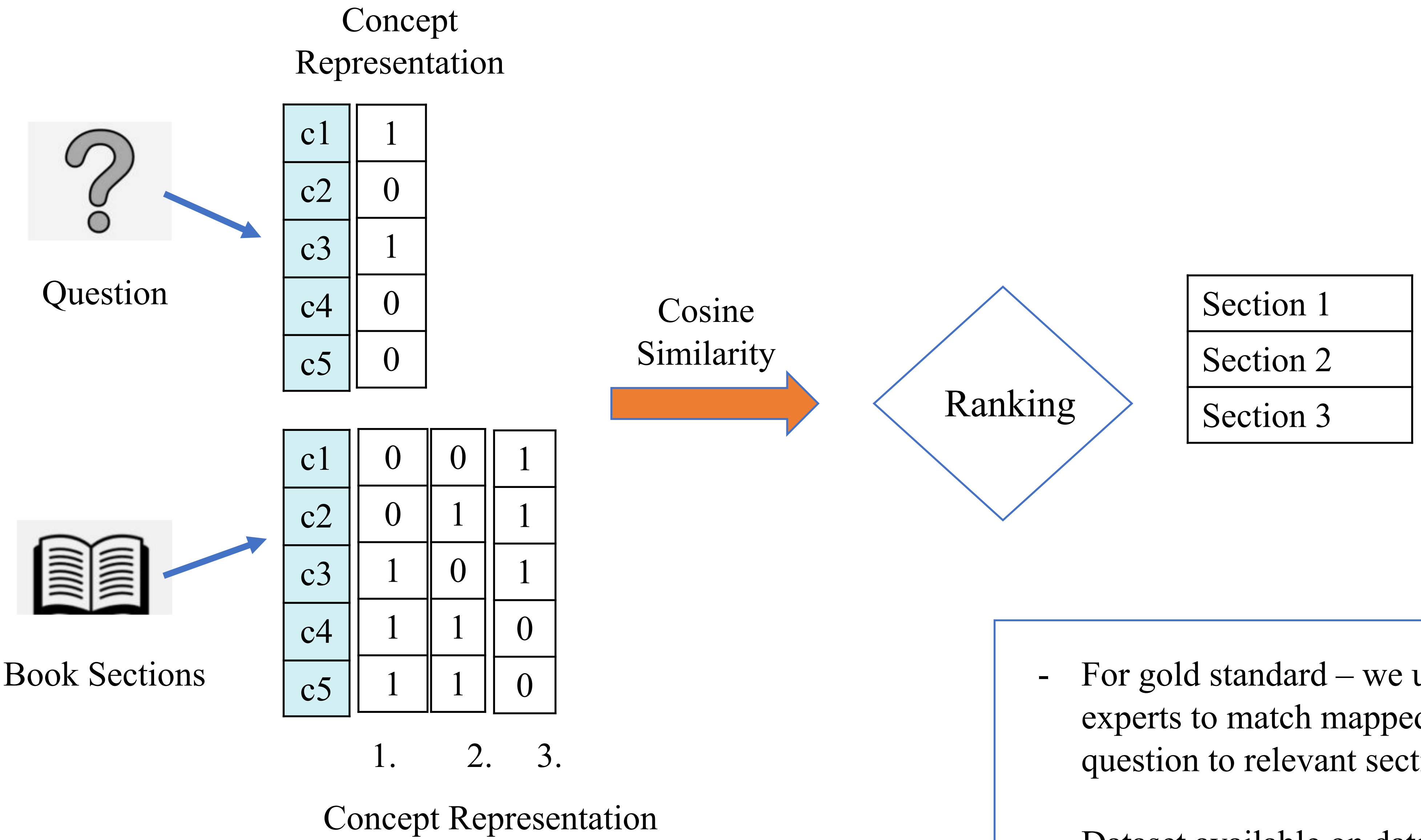


ReadingCircle: J. Guerra, D. Parra, and P. Brusilovsky. Encouraging online student reading with social visualization. In CEUR Workhop, AIED 2013

Remedial Recommendation



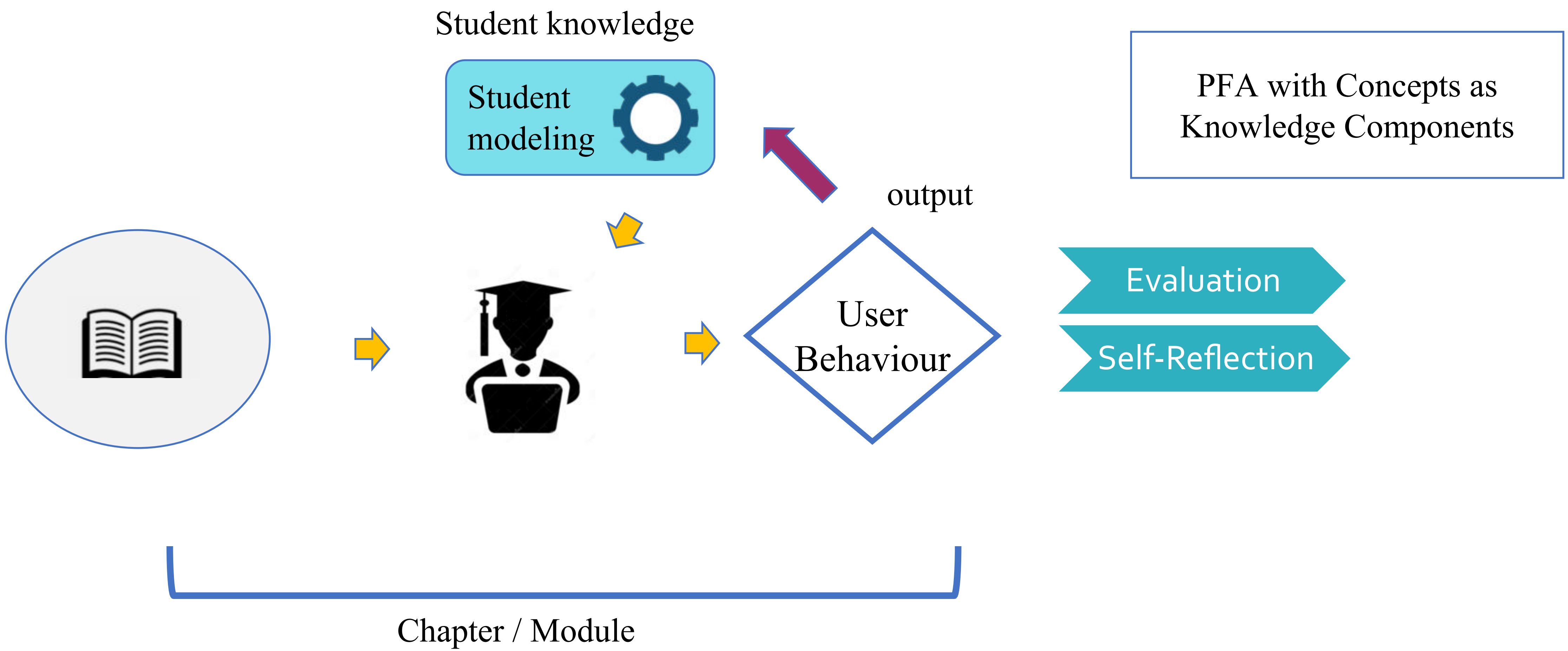
Concept-based Remedial Recommendation



- For gold standard – we used experts to match mapped each question to relevant section.
- Dataset available on datashop[1]

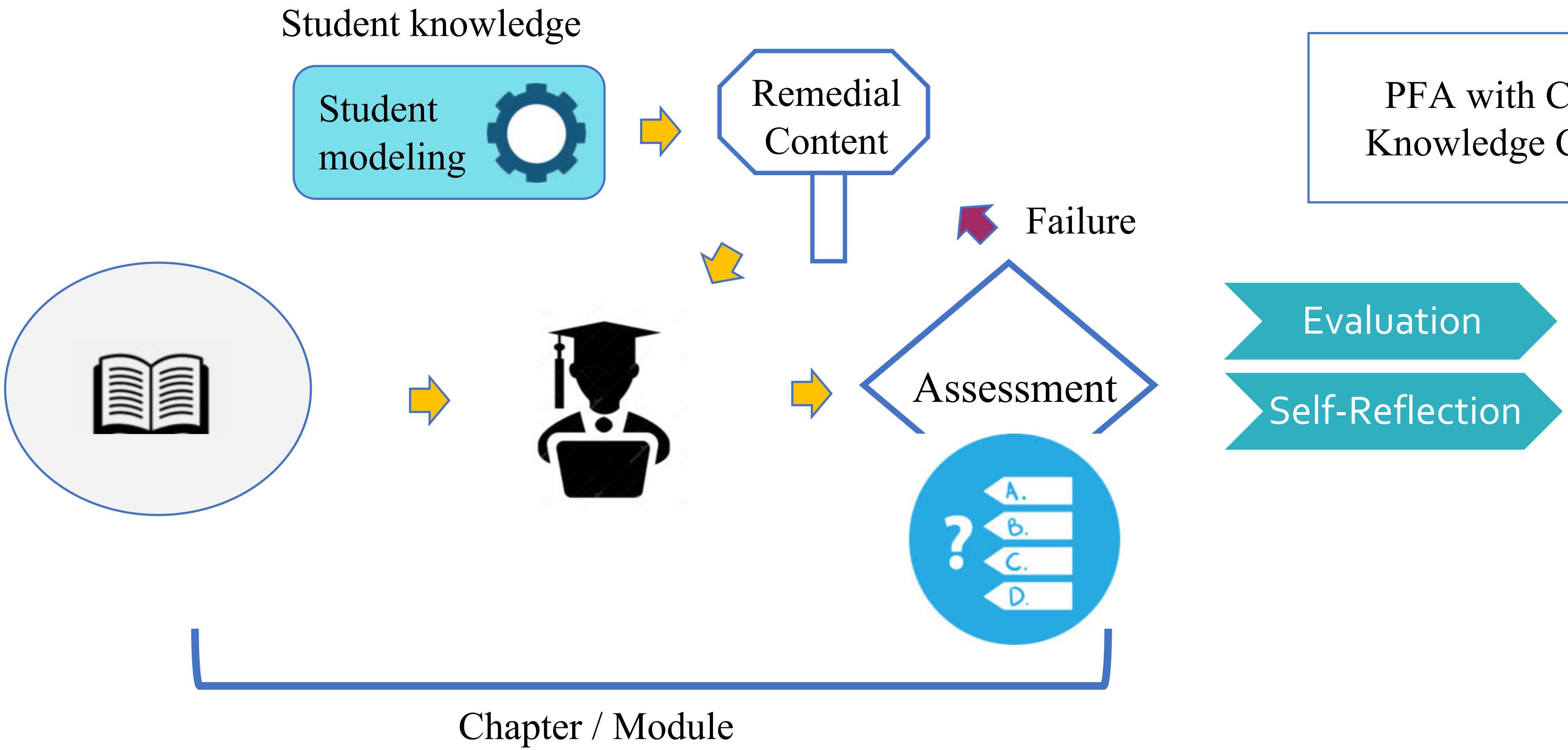
[1] https://pslcdatashop.web.cmu.edu/Project?id=637

Knowledge based Remedial Recommendation



PFA – Performance Factor Analysis

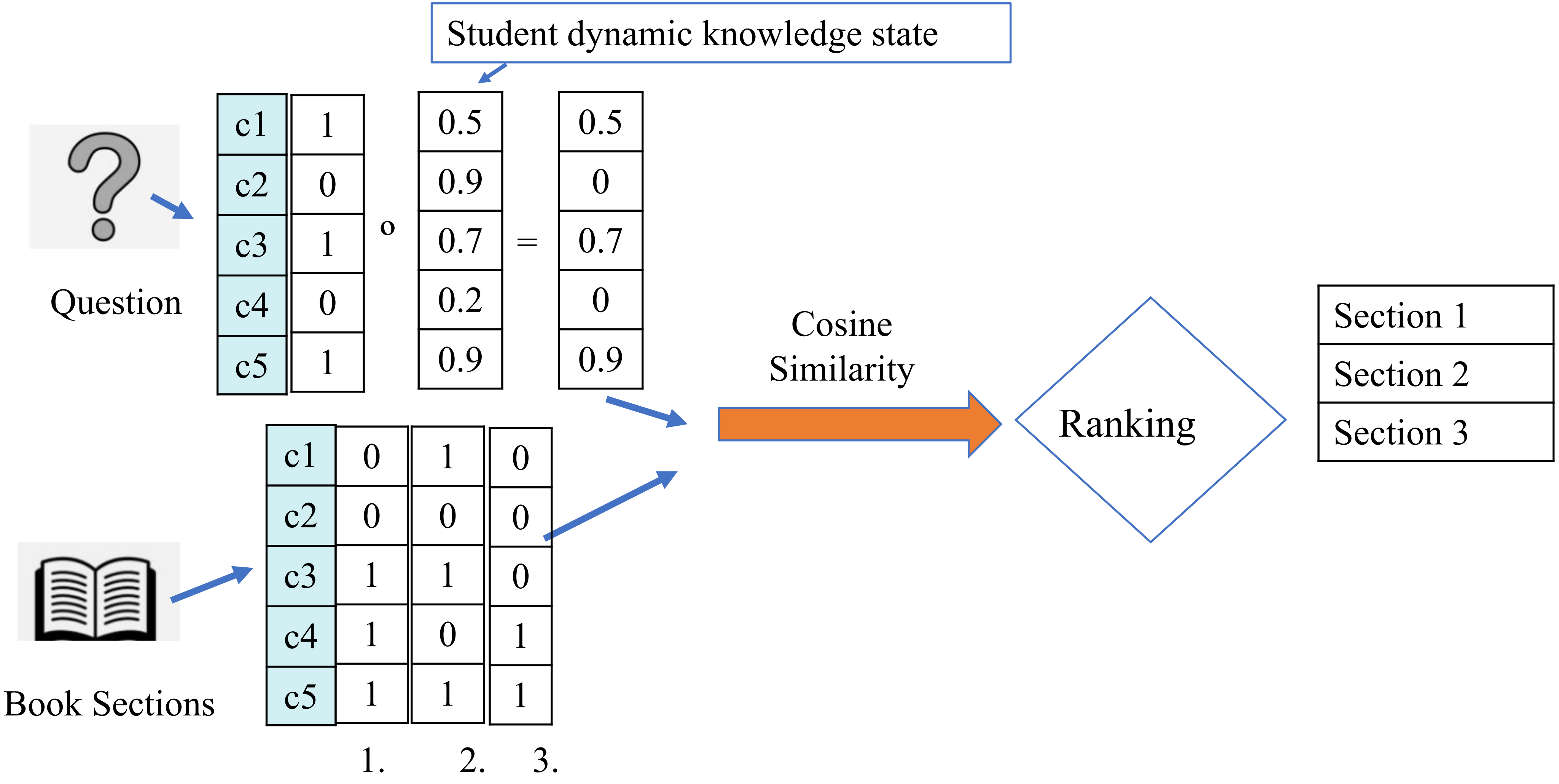
Knowledge based Remedial Recommendation



PFA with Concepts as Knowledge Components

PFA – Performance Factor Analysis

Knowledge-based Remedial Recommendation



[1] https://pslcdatashop.web.cmu.edu/Project?id=637

Results: Keyphrases as Concepts

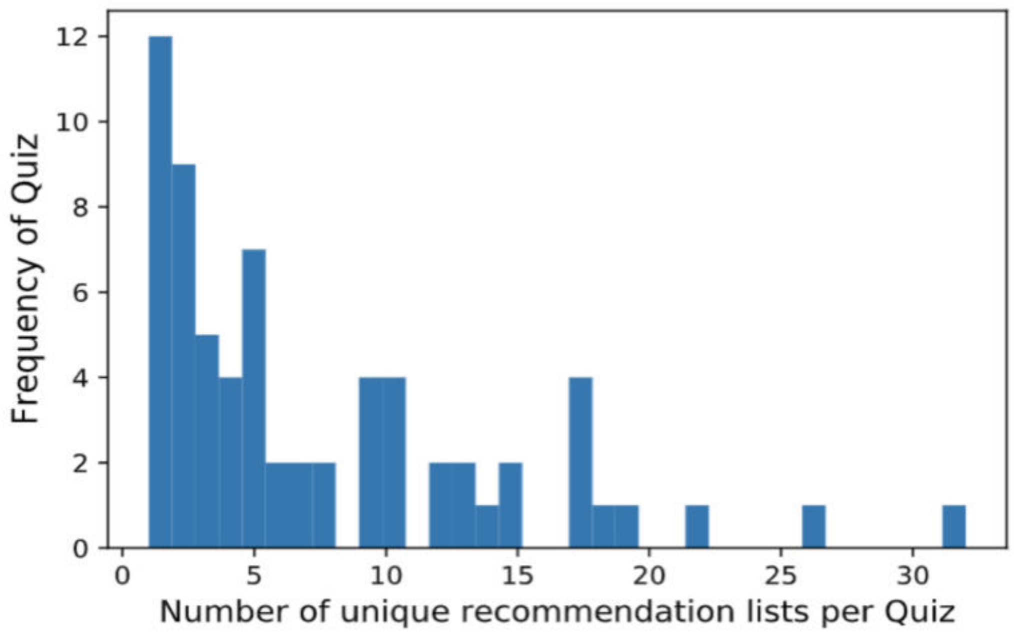
	Text-based	Concept-based	Knowledge-based
		MAP@5	MAP@5
Text Similarity	0.74	-	-
Expert Concepts	-	0.8618	0.8390
TextRank	-	0.8314	0.8397
CopyRNN	-	0.8466	0.8405
TopicRank		**0.8990***	**0.8885***

All are better then text-based

Bold indicates significantly better than baseline with expert generated concepts

Knowledge-based Remedial Recommendation

It should recommend different thing to different people ?



- For 86% of quizzes students are recommended different sections
- There are cases where a failure on a quiz is mapped to more than 15 unique sections
- This shows that student knowledge does provide personalized recommendations

Takeaway messages:

- Students' learning can be helped with knowledge units identified from text
- Keyphrases can act as useful knowledge units in supporting students' learning

Outline

- Basics of Keyphrases: Definitions and Importance
- Identification of Keyphrases: Extraction and Generation
- Applications of Keyphrases: knowledge unit for supporting student learning
- Applications of Keyphrases: knowledge unit for recognizing patients’ concerns
- Applications of Keyphrases: knowledge unit for interactive machine learning
- Conclusions

Weakly Supervised Medical Entity Extraction and Linking for Chief Complaints

Zhimeng Luo, Zhendong Wang, Rui Meng, Diyang Xue, Adam Frisch and Daqing He

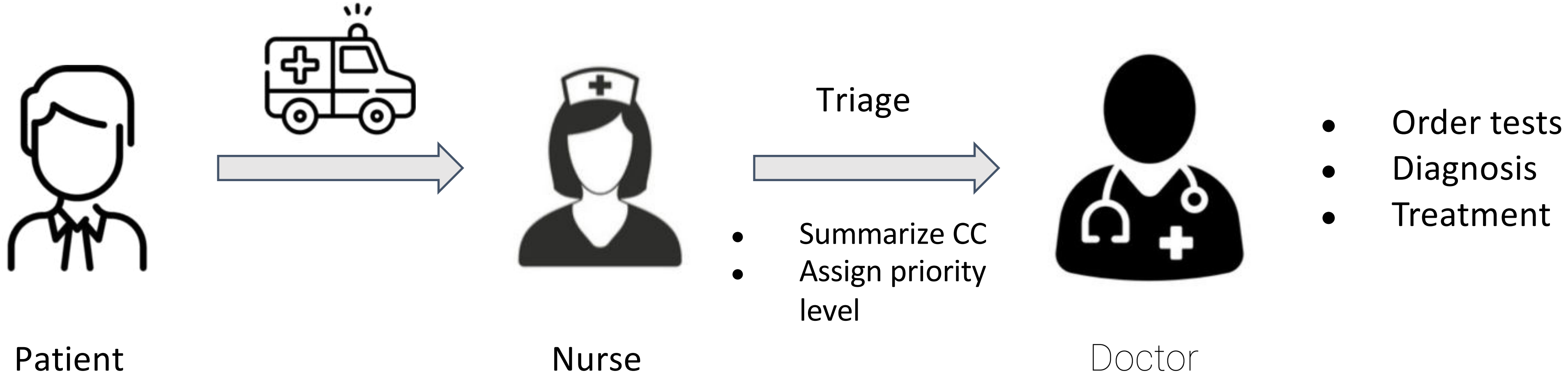
Chief Complaints

The beginning of physician's diagnosis process in emergency department (ED) is guided by the patient's chief complaint (CC).

Chief Complaint is a record to summarize:

- reason for encounter
- current symptoms
- medical history

Chief Complaint:
migraine with neck/back pain, fever



Chief Complaints

*Chief complaint **record/instance***: ha light headed fatigue r arm pain

*Chief complaint **entity mentions***:
(***Span***: location of each mention)
ha light headed fatigue r arm pain

*Chief complaint **entity concepts***:
(from HaPPy's ontology)
headache ; dizziness ; fatigue ; arm pain

Characteristics of chief complaint:

- Short free-text descriptions, with large variation (abbreviaions, synonyms, ...)
- A record may contain one or multiple CC entity mentions
- Span of each concept is important

  Doctors want to know span along with the concepts

  To summarize entity variants to improve the existing ontology

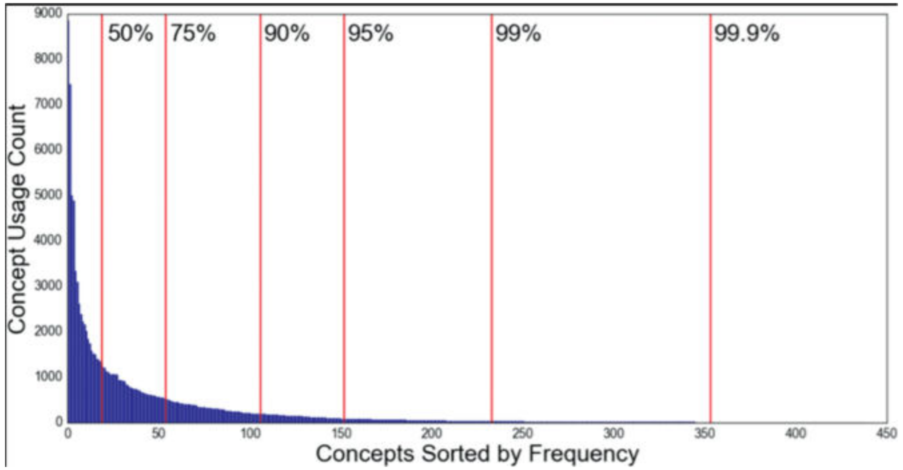
HaPPy ontology

- The first publicly available large-scale CC ontology
- Containing 692 unique concepts, 2,118 synonyms, and 30,613 descriptions.

- We found that
  - Direct match result in low matching recall following HaPPy's instruction across health care system (on UPMC corpus).
  - automatic CC extraction and linking is necessary
  - use it as ground truth concept labels in our project

Table 4
A typical presenting problem, synonyms, and nonvisible descriptions

Presenting problem ($n = 1$)	Displayed synonyms ($n = 4$)	Nonvisible descriptions ($n = 516$)
Headache	Headache B Headache L Headache R Headache	H/A HA HEAD ACHE HEADACHE HEADACHES HEAD PAIN MIGRAINE ...(more)... (L) HEAD ACHE (L) HEAD PAIN ...(more)... RT-SIDED HEADACHES RT-SIDED MIGRAINE



Horng, S., Greenbaum, N. R., Nathanson, L. A., McClay, J. C., Goss, F. R., & Nielson, J. A. (2019). Consensus Development of a Modern Ontology of Emergency Department Presenting Problems-The Hierarchical Presenting Problem Ontology (HaPPy). Applied clinical informatics, 10(3), 409–420. https://doi.org/10.1055/s-0039-1691842

Limitations of Related Work

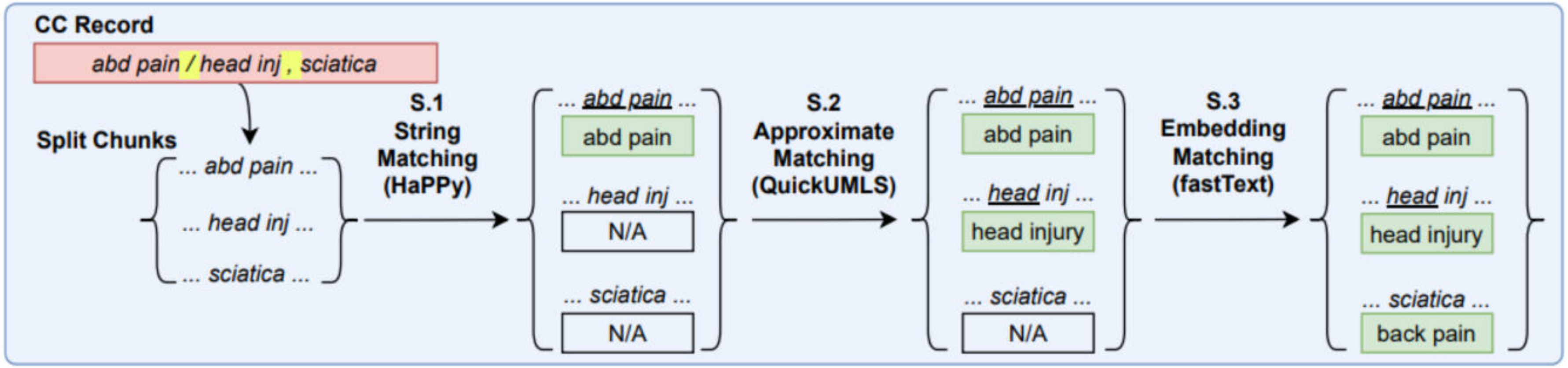
Limitations:

- Cannot identify **multiple** CC concepts within a record
- Cannot identify the **span** of each CC entity mentions
- Lack of large-scale annotations **containing span information**

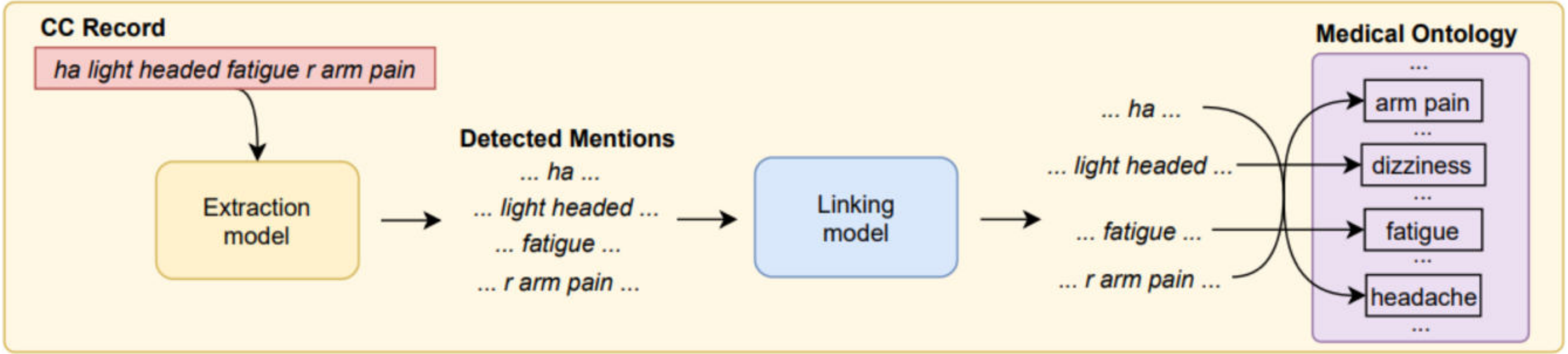
We propose to view the task as:

- **entity extraction**
    - identifying the actual mention span of each entity in a free text
- **entity linking**
    - linking each entity mention to a concept in a CC ontology
- **Weak supervision**
    - Identifying noisy patterns in text without manual annotations?

Proposed method: WeSEEL



(a) Process flow of weak label generation. Three examples are shown and successfully matched to concepts in the ontology at different stages (indicated in green box).

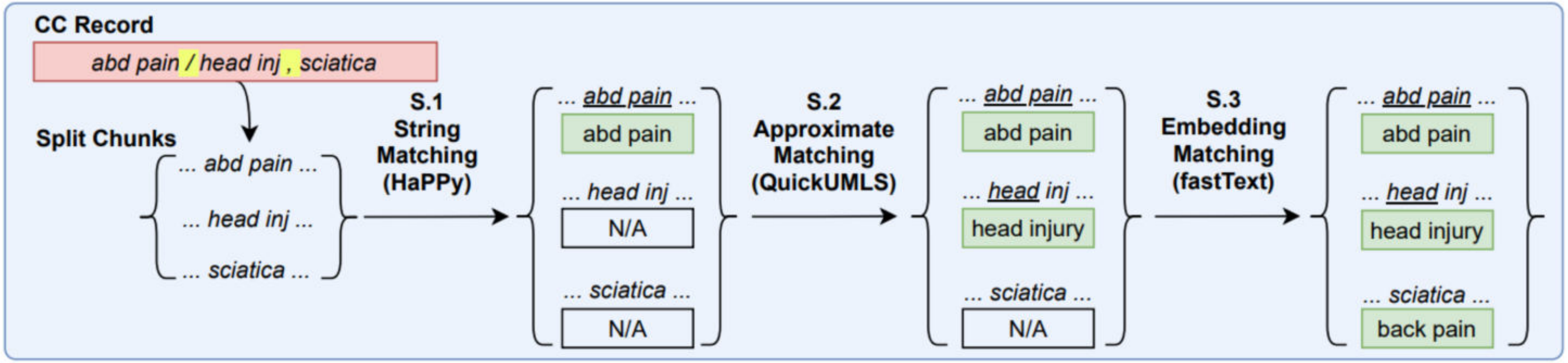


(b) Schematic diagram of our model for entity extraction and linking in chief complaint.

Figure 1: Overview of our proposed method WeSEEL (Weakly Supervised Entity Extraction and Linking in chief complaint).

Weak label generation

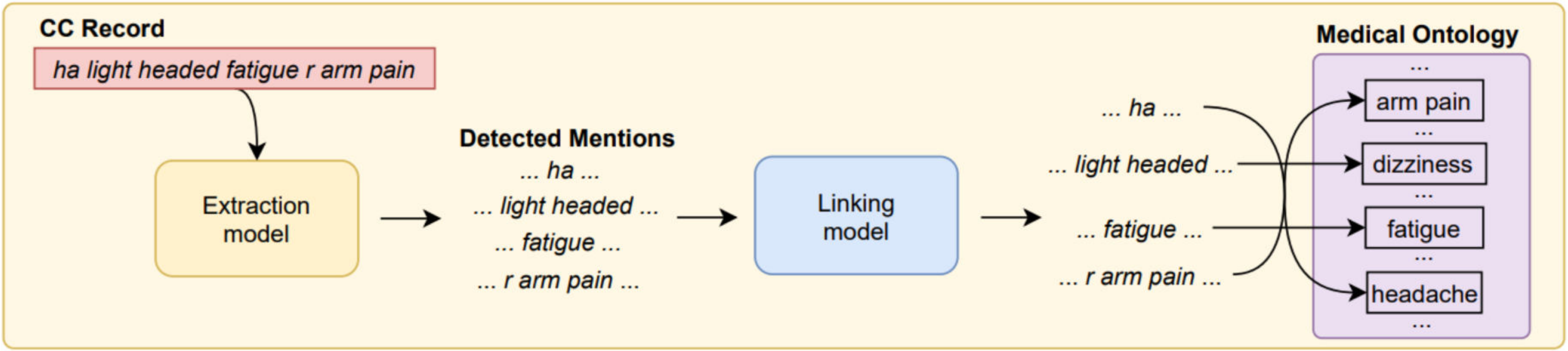
To assign concept labels with corresponding span information for model training, a **Split-and-Match** algorithm is proposed as follows:



(a) Process flow of weak label generation. Three examples are shown and successfully matched to concepts in the ontology at different stages (indicated in green box).

- **Exact** string matching (HaPPy ontology)
- **Approximate** string matching (QuickUMLS)
  - resolve misspelling and lexical variations
- **Embedding**-based matching (fastText)
  - enables semantic matching

Two-step model



(b) Schematic diagram of our model for entity extraction and linking in chief complaint.

- Entity Mention **Extraction** (extraction model)
- **Linking** Entities to Ontology (linking model)

Entity Mention Extraction

Formulate as a **sequence labeling** problem that follows the BIO tagging:

- e.g., "10 wks/n/v/d" -> "10 wks / n / v / d"

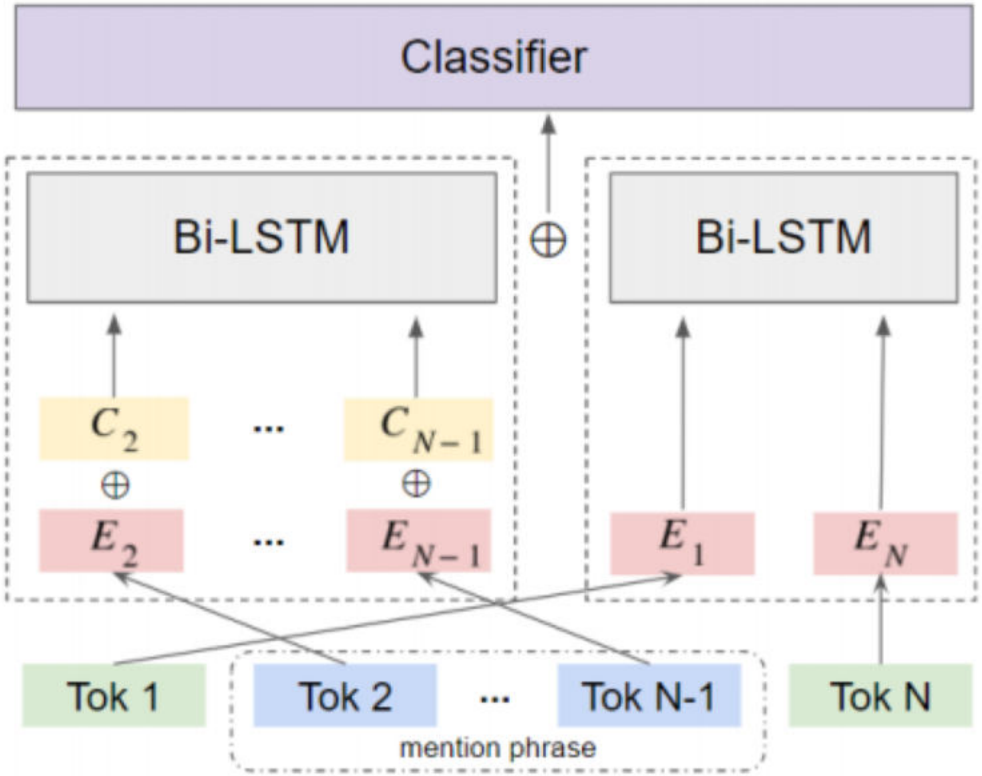
  B I O B O B O B

- BERT token classification model (CCME)
- Soft label
  - adjust the label smoothing to accommodate the weak span labels
  - For each word in a chunk, set the probability of a weak target label as the **similarity** between a chunk and its corresponding ontology concept

Tokens	Adjusted Token Weight	Adjusted Target Vector (B/I/O)
chest pain and	0.9	[0.9, 0.0, 0.1]
chest pain and	0.9	[0.0, 0.9, 0.1]
chest pain and	0.0	[0.0, 0.0, 1.0]

Linking Entities to Ontology

- Link each entity mention extracted from CC records to a given ontology through a classification model.
- Use BiLSTM as basic layer.
- Propose two additional input embeddings along with the major mention word embedding:
  - Surrounding context embedding
    - to consider context information
  - Character embedding of mention tokens
    - to consider lexicon variations



Architecture of the model for entity linking in chief complaints (CCEL)

Data set

- 1,232,899 free-text CC records were collected from UPMC Health Service system
- covering the period of 2015 to 2017 from 15 hospitals.
- All EDs use the same electronic health record system, but do not mandate a specific data entry format.
- A test set of 1,013 random instances was annotated by an ED clinician.
- HaPPy ontology was selected as the target ontology to link entities
    - shrunk the original ontology (692 to 501 concepts) by removing child nodes that have no significant clinical difference with their parent node
    - e.g., “ruq abdominal pain” (“ruq” means right upper quadrant) were folded to parent node “abd pain”
- All data have been checked by IRB

Data set

No. cc	instances
0	61
1	388
2	371
3	143
4	39
5	10
6	1

No. concepts in each record in test set.

Punctuation	#Tokens	Conditions	#Instances	Percentage	#Sampled
w/	Single	Only contains "/"	7,739	-	16
w/	Single	Only contains ","	1,285	-	3
w/	Single	Only contains "\.br\"	529	-	1
w/	Single	Only contains "..."	459	-	1
w/	Single	Others	1,097	-	2
w/	Single	-	11,109	2.30%	23
w/	Multiple	Only contains ","	154,457	32.01%	324
w/	Multiple	Only contains "/"	78,129	16.19%	164
w/	Multiple	Only contains "..."	4,017	-	8
w/	Multiple	Only contains "\.br\"	3,248	-	7
w/	Multiple	Only contains "+"	2,154	-	5
w/	Multiple	Only contains ":"	1,902	-	4
w/	Multiple	Only contains "."	1,895	-	8
w/	Multiple	Only contains "\"	1,695	-	4
w/	Multiple	Only contains "&"	1,599	-	3
w/	Multiple	Contains "," and "/"	17,555	3.64%	37
w/	Multiple	Others	11,182	2.32%	23
w/	Multiple	-	277,833	57.58%	583
w/o	Single	Others	5,108	-	11
w/o	Multiple	Only contains "and"	19,078	3.95%	40
w/o	Multiple	Only contains "at"	1,292	-	3
w/o	Multiple	Others	168,118	34.84%	353
w/o	-	-	193,596	40.12%	407

Table 3: Distribution of unique instances with conditions. The test set is sampled based on the distribution of conditions, shown as #Sampled. The percentage less then 2% are not shown in the table.

Evaluation metrics

- We adopt the evaluation protocol of SemEval 2013 task 9.1
- Mention extraction
  - "Partial" mode (partial boundary match, regardless of the type)
  - "Exact" mode (exact boundary match, regardless of the type)
- Entity linking
  - "Entity type" mode (partial boundary match and correct entity type)

Test results of entity extraction

Models	Partial Match Precision	Partial Match Recall	Partial Match F1	Exact Match Precision	Exact Match Recall	Exact Match F1
S&M (HaPPy)	95.81	36.90	53.28	92.69	35.70	51.54
S&M (QuickUMLS)	81.15	46.04	58.75	67.46	38.28	48.84
S&M (Embedding)	69.64	**57.36**	62.91	55.66	45.84	50.27
CCME-LSTM	78.45	48.26	59.76	67.25	42.39	52.00
CCME-BERT	81.37	53.43	64.50	71.46	48.52	57.80
CCME-BERT (soft)	83.41	56.70	**67.51**	72.95	**49.59**	59.04
CCME-ClinicalBERT (soft)	83.35	56.46	67.32	72.94	49.41	58.91
CCME-BERT (soft) + S&M (HaPPy)	**96.28**	44.86	61.20	**92.94**	43.30	**59.08**
CCME-BERT (soft) + S&M (QuickUMLS)	86.13	51.82	64.71	76.86	46.24	57.74

Table 4: Entity extraction performance of different models. Scores are computed in **Partial Match** and **Exact Match** mode of SemEval'13. The best/2nd-best scores in each column are in bold/underlined.

- Most neural models outperform the matching-based methods, indicating that machine learning models can learn task-relevant inductive bias from weak labels

Test results of entity linking

#	Extraction Model	Linking Model	Precision	Recall	F1
s.1	†S&M (S.1)	†S&M (S.1)	**98.02**	37.75	54.51
s.2	†S&M (S.1 + S.2)	†S&M (S.1 + S.2)	79.34	45.02	57.44
s.3	†S&M (S.1 + S.2 + S.3)	†S&M (S.1 + S.2 + S.3)	57.73	47.54	52.14
n.1	CCME-BERT (soft)	fastText (single-label)	92.93	21.11	34.41
n.2	CCME-BERT (soft)	fastText	89.27	26.69	41.09
n.3	CCME-BERT (soft)	BERT	83.48	34.52	48.84
n.4	CCME-BERT (soft)	CCEL	84.36	45.22	58.88
b.1	QuickUMLS	MedType (EHR)	53.32	23.09	32.23
m.1	†S&M (S.1 + S.2)	fastText (single-label)	89.07	15.77	26.79
m.2	†S&M (S.1 + S.2)	fastText	86.49	19.52	31.85
m.3	CCME-BERT (soft)	†S&M (S.1)	<u>97.86</u>	45.60	62.20
m.4	CCME-BERT (soft)	†S&M (S.1 + S.2)	85.64	<u>51.53</u>	<u>64.34</u>
m.5	CCME-BERT (soft)	†S&M (S.1 + S.2) + CCEL	86.28	**55.43**	**67.49**

Table 5: Entity linking performance. Scores are computed in **Entity Type** mode of SemEval'13. The **best**/<u>2nd-best</u> scores in each column are in bold/underlined. †S.1, S.2, S.3 refer to string matching algorithms in Figure 1(a).

- CCME-BERT models are good at identifying entity mentions, while matching methods are good at classifying concepts given identified mentions

Effect of Weak Supervision

We simulate a fully supervised setting: 80% training, 20% testing from annotated test set (report average scores from 5-fold CV)

- Supervised: training models with annotated data only;
- Fine-tuning: pre-train using weak labels and fine-tune it with the annotated data.

Training	P	R	F1
WeakSup	**83.41**	56.70	67.51
Supervised	77.76	**89.66**	83.29
Fine-tuning	82.25	85.98	**84.07**

Table 7: Extraction performance (Partial Match) of CCME-BERT (soft) with three training strategies.

- Trained with little annotated data, CCME-BERT achieve decent results on mention extraction
- Pre-training the model with weak labels can be beneficial

EHR: admission note

- EHR Notes contains narrative information about a patient’s current and past medical history.
    - Many types of notes in the EHR including: admission notes, assessments, SOAP notes, exams, reports, and etc.
- **Admission notes** document the reasons why a patient is being admitted to a hospital or other facility, the patient's baseline status, and the initial instructions for that patient's care. Its important components are
    - Chief Complaints (CC): "abdominal pain"
    - History of Present Illness (HPI): "Pt is a 30 yo female (with a PMH of x and y) presenting with a 3 hour history of abdominal pain..."
    - Review of Systems (ROS): “immunologic : negative. \n musculoskeletal : right lower extremity pain and swelling.. ”
    - Medical Decision Making (assessment): similar to the first line of the HPI, but with a greater emphasis on clinical reasoning.
    - Diagnosis and Plan

Examples note 1

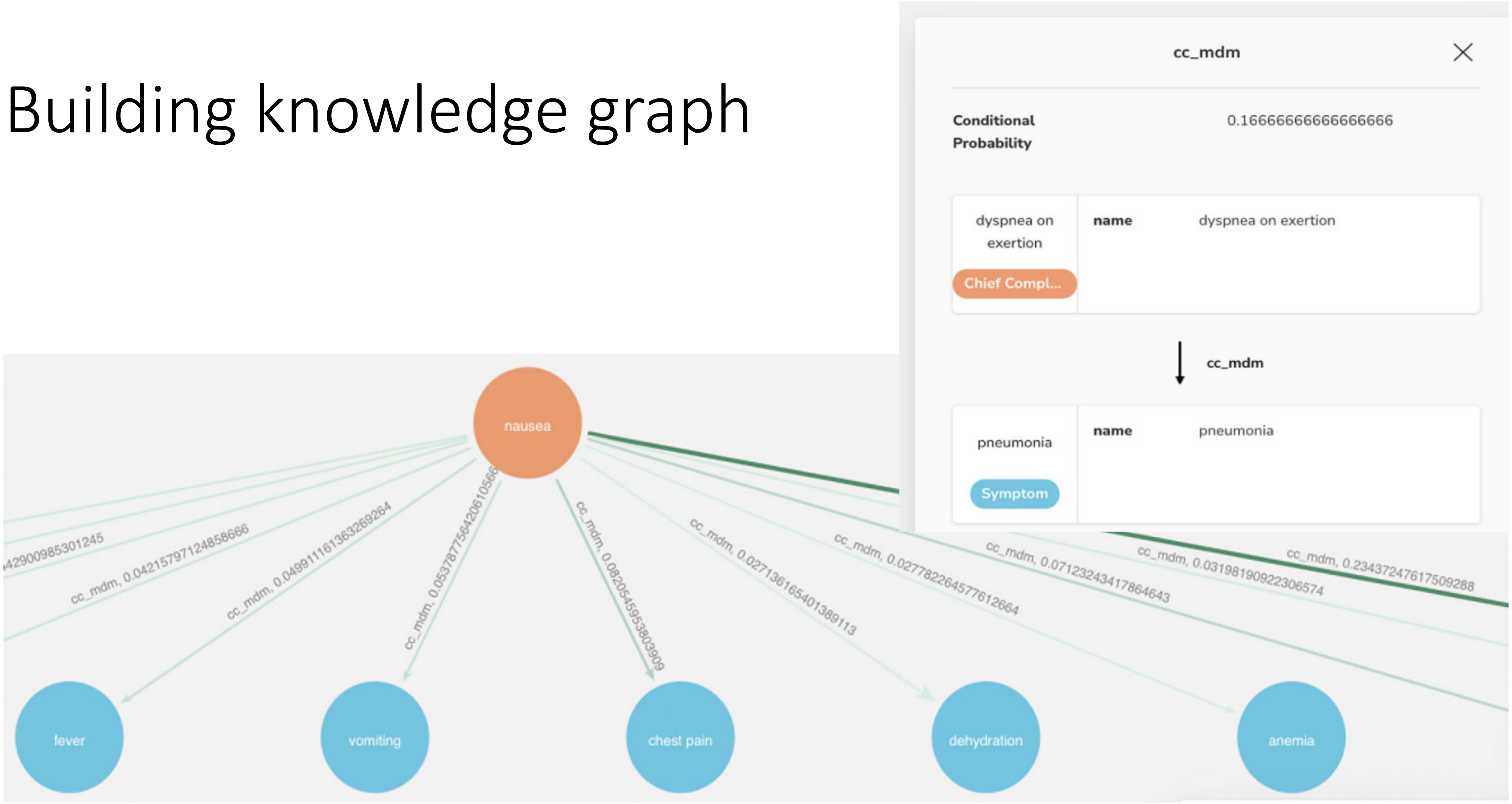
- CC section
  - "easy bruising (ecchymosis), rle pain (leg pain) and swelling (leg swelling)"
- ROS section
  - "musculoskeletal : right lower extremity pain (leg pain) and swelling (leg swelling)"
  - "integumentary : petechiae"
- HPI section
  - "patient is a 39 y/o female with pmh +lupus anticoagulant c/b recurrent pes now on warfarin, kidney stones, asthma (asthma exacerbation), seizure disorder, bipolar disorder, hypothyroidism who presents for evaluation of right lower extremity pain (leg pain) and easy bruising (ecchymosis). patient reports recent fall on saturday 5/6/17, tells me that her shoes "caught up" and then she tripped and fell to the ground hitting her back and her right lower leg. denies trauma to the head or loss of consciousness…"
- Medical Decision Making section
  - "patient presents for evaluation of increasing ecchymosis, right leg pain and swelling suspect this is primarily related to significant hematoma in the right leg though there is concern for dvt (prior to knowledge of her inr). right lobe chevy dopplers obtained shows no evidence of dvt. her inr was found to be significantly supratherapeutic…because of pain control, need to trend hemoglobin given the possibility of significant bleeding into the right leg.."
- Diagnosis
  - "supratherapeutic inr, right lower extremity hematoma, right leg swelling

CCs

Symptoms

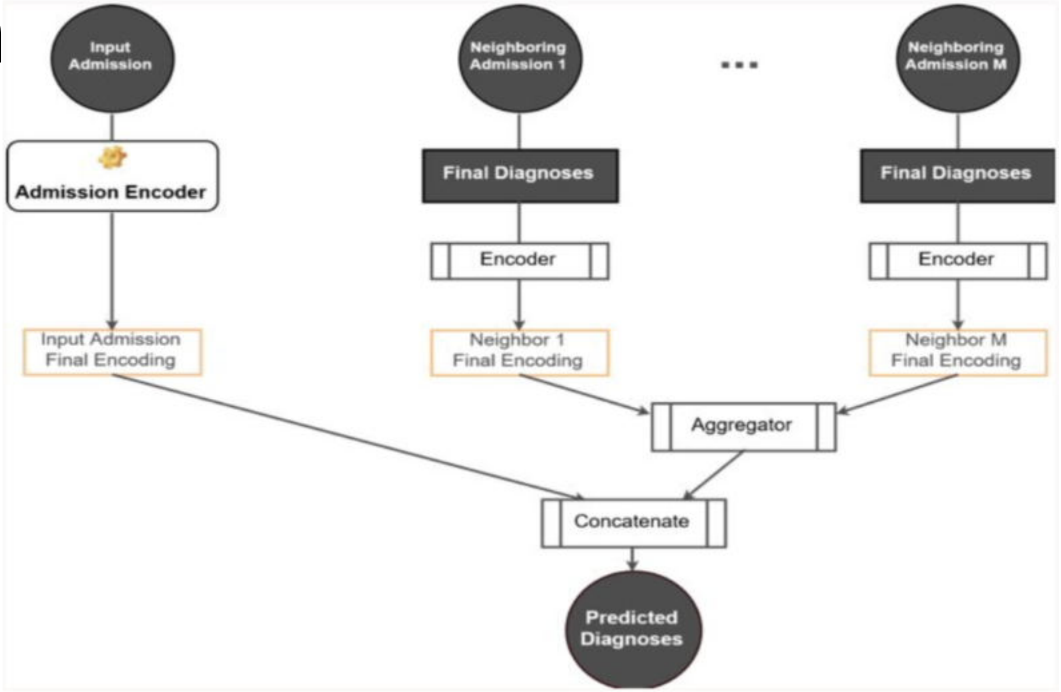
Diagnosis

Building knowledge graph



Showcase the usage of the graph

Neighbor Extraction



Diagnosis Prediction

Takeaway messages:

- Chief complaints are noisy yet important keyphrases in clinical text
- Our method paves a foundation for further exploration of clinical text using CC as knowledge units

https://blogs.oracle.com/ai-and-datascience/post/graph-machine-learning-for-enhanced-healthcare-services

Outline

- Basics of Keyphrases: Definitions and Importance
- Identification of Keyphrases: Extraction and Generation
- Applications of Keyphrases: knowledge unit for supporting student learning
- Applications of Keyphrases: knowledge unit for recognizing patients' concerns
- Applications of Keyphrases: knowledge unit for interactive machine learning
- Conclusions

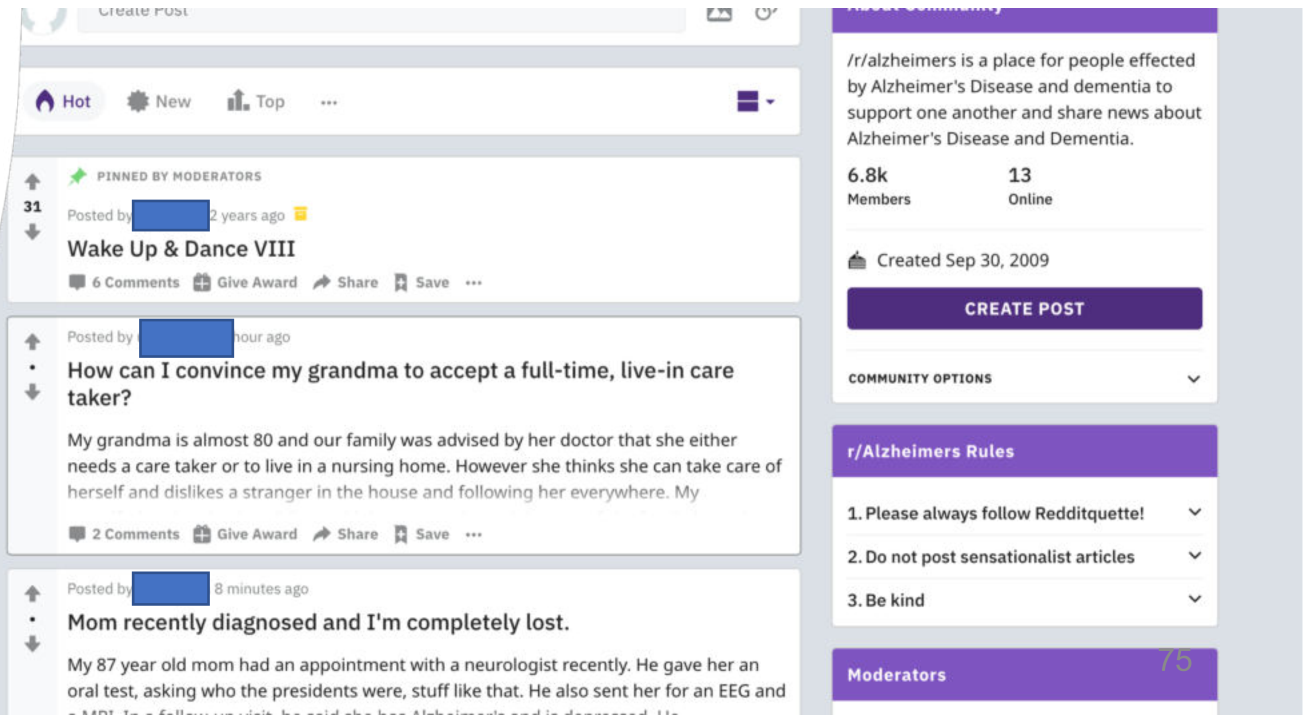
Characterizing Dementia Caregivers' Information Exchange on Social Media: Exploring an Expert-Machine Co-Development Process Zhendong Wang[1], Ning Zou[1], Bo Xie[2], Zhimeng Luo[1], Daqing He[1], Robin C. Hilsabeck[2], Alyssa Aguirre[2]

Background

- Alzheimer's disease and related dementias (ADRD) are a major public health concern
- In the U.S., about 5.6 million Americans age 65 and over were living with ADRD in 2019[1]
- Caregiving for people with ADRD is stressful[2][3][4]

Background

- Social media platforms have introduced novel mechanisms supporting online health information seeking and sharing
- Research on ADRD caregivers' information exchange via social media platforms remains limited

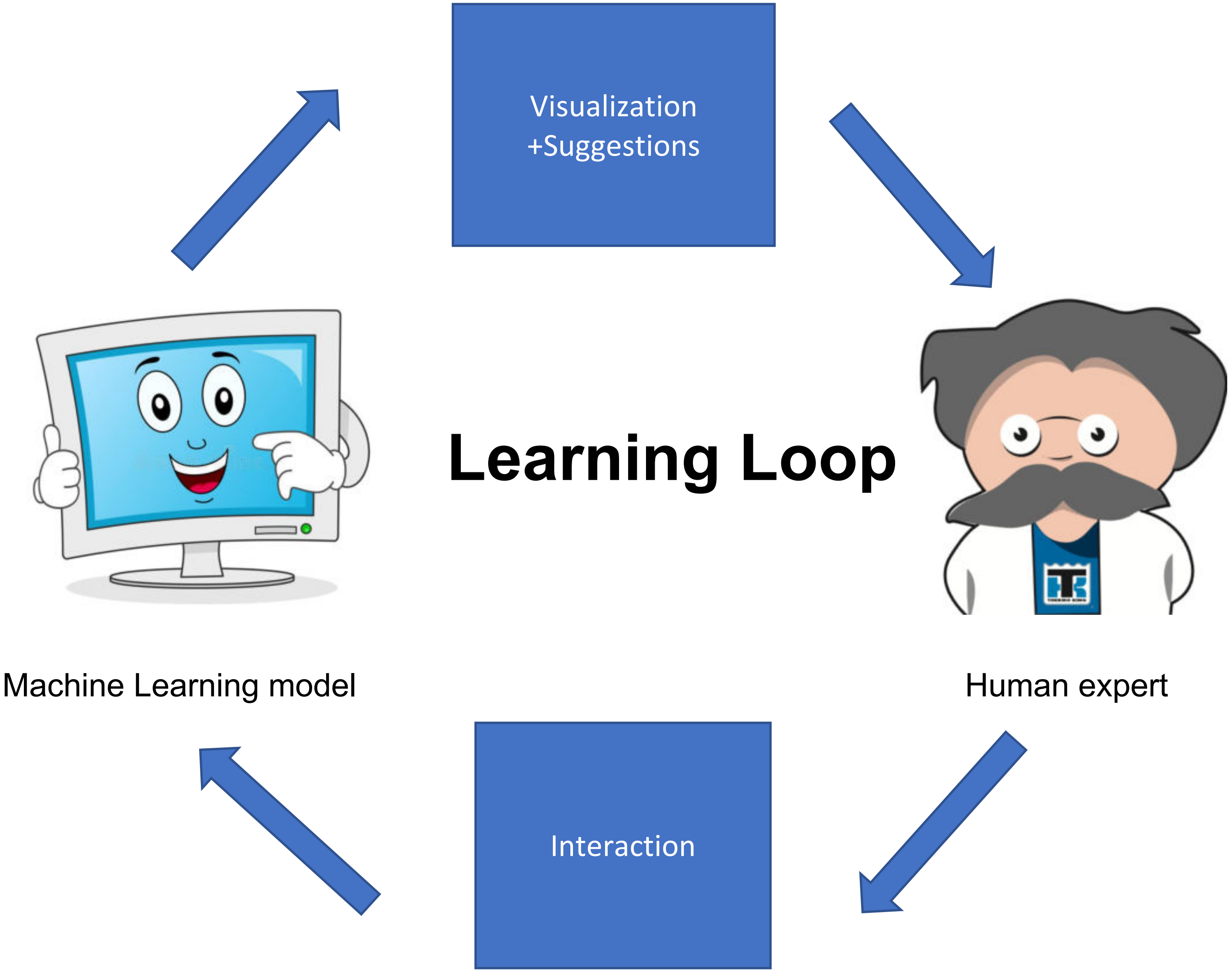


Previous works

- Expert Analysis: relies on human experts to manually analyze social media content
  - Accurate but time consuming, costly, and problematic for large amounts of rapidly growing social media data.
  - Social media is also new for human experts
- Automatic Exploration: uses machine learning or text mining algorithms
  - Able to overcome these limitations of expert analysis
  - Requires large annotated data from human experts
  - Lacks iterative interaction or knowledge exchange between human experts and automatic algorithms

Previous works

- Interactive Machine Learning (IML):
  - IML utilizes learning loop with interaction from human experts to iteratively increase performance of machine learning model with less human efforts
  - IML has been applied successfully in wide range of domains[5][6][7] but not yet in health information exchange.



Proposed method: EMC Process

- Expert-Machine Co-development (EMC) Process:
  - Create a Health Information Want(HIW) framework to analyze the category and keywords of ADRD online posts
  - IML based interactive process with rich interactions
  - maximize the strengths of both human experts and automatic algorithms
  - Minimize human efforts
- Components:
  - Component 1: Expert Analysis of ADRD Caregivers’ Information Exchange
  - Component 2: Automatic Exploration (AE) of ADRD Caregivers’ Information Behaviors

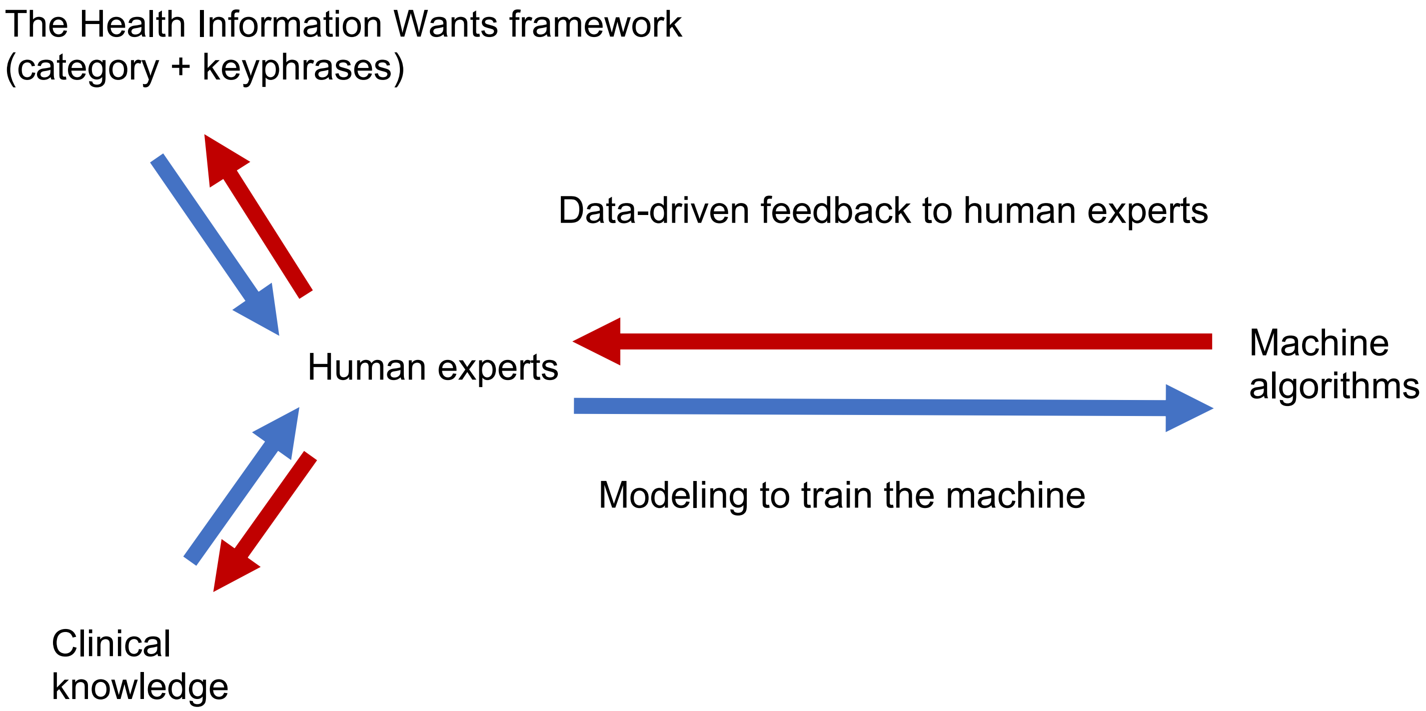


Figure 1. EMC Process

Research Aims

- Aim 1: What ADRD-related information do caregivers exchange on social media?
- Aim 2: How an interactive learning system can be designed to enable the EMC process
- Aim 2: What roles can keyphrases extracted from online posts can play to help both human experts and the machine learning system?

Component 1: Expert Analysis of ADRD Caregivers Information Exchange

- Goal: create an initial framework for ADRD information exchange analysis
- Collecting Data: 823 posts from reddit group of Alzheimer
- Health Information Wants framework (HIW-ADRD) development:
  - 7 Categories
  - 176 keyphrases
  - 200 manually annotated posts

Type of information	Sample keywords	Number(%) of posts
Treatment / Medication/ Prevention	Drug; oriental; acupuncture; vitamins	8(4)
Characteristics of / Experience with the health condition/ Diagnostic procedures	Diagnosis; complication; cause; prognosis; process; symptom; memory loss; lab test; MRI; PET; blood	17(8.5)
Daily care for a patient at home / Care for a caregiver (practical strategies or tips, not psychosocial)	Wandering; bath; hygiene; sleep; eat; driving	28(14)
Care transition and coordination / End-of-life care (practical, not psychosocial)	Adult day care; rehab; hospital; memory care; nursing home; hospice	13(6.5)
Psychosocial aspects	Stress; lonely; heartbreaking; overwhelmed; venting	66(33)
Resources/Advocacy/Scientific updates/Research participation	Lobby; fundraising; clinical trial; news; article; scientist	63(31.5)
Legal/Financial/Insurance	Power of attorney; POA; living will; Medicare; Medicaid	5(2.5)

Table 1. The HIW-ADRD 3.0 framework

Component 2: Automatic Exploration (AE) of ADRD Caregivers' Information Behaviors

- Goal: Improve the HIW-ADRD framework
  - Tuning existing keyhrases
  - Help human experts discover better keyphrases
  - Improve the accuracy of model

- AE Process
  - Initial Model training
  - Interactive Learning loop
    - Keyphrases tuning recommendation
    - AE assisted exploration and feedback
    - Annotation and stop recommendations

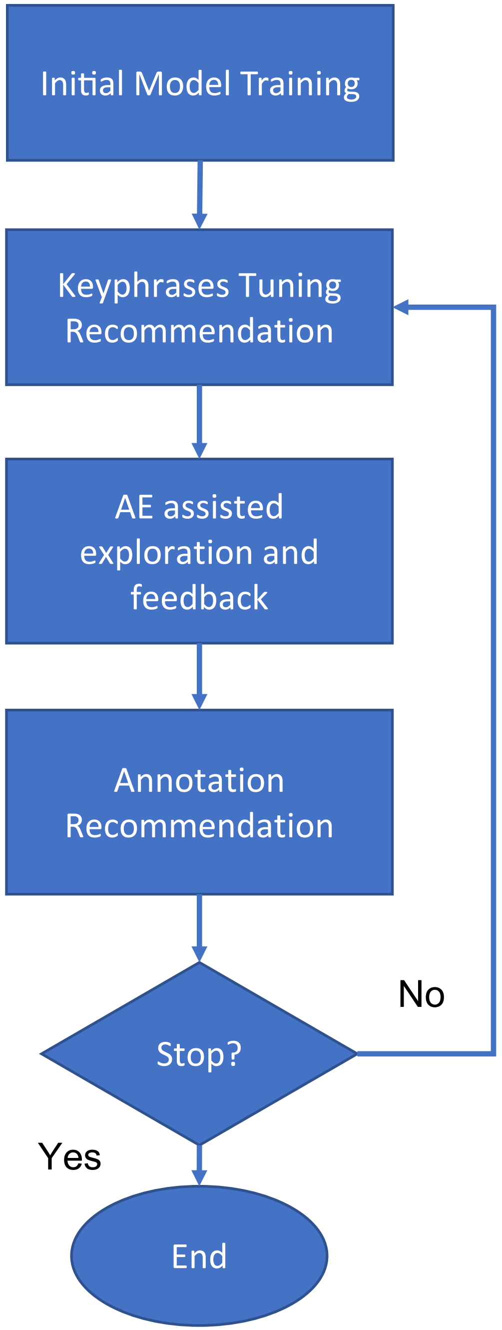


Figure 2. Automatic Exploration Process

Component 2.1: Initial Model training

- Model
  - Input: document representations of keyphrases and their frequency
  - Output: HIW-ADRD categories
- Initial training dataset: annotation from result of Component 1

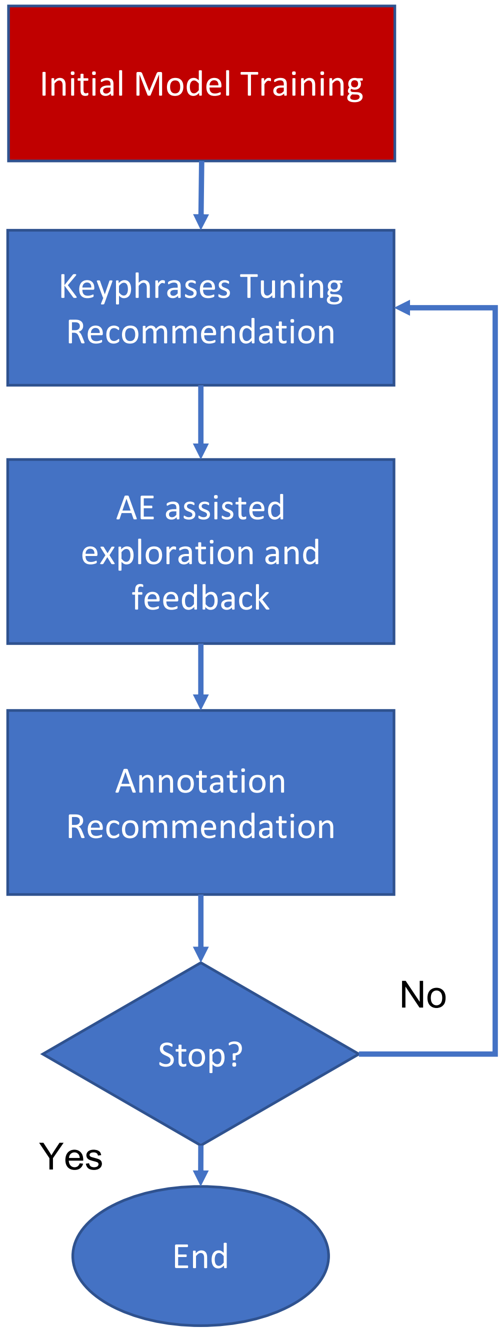


Figure 2. Automatic Exploration Process

Component 2.2: Keyphrases tuning recommendation

- recommendation criteria:
    - Mutual Information (MI) Score (normalized) between keyphrases and category
    - Importance(I) Score between keyphrases and model
    - Keyphrase Frequency (KF) for keyphrases in posts
- 4 Keyphrases Tuning(KT)
    - potentially good (PG): high MI, low I
    - potentially bad (PB): high I, low MI
    - low frequent(LF): too small KF
    - New Keywords (NK): Not in the existing framework, but has potentially high MI and enough KF

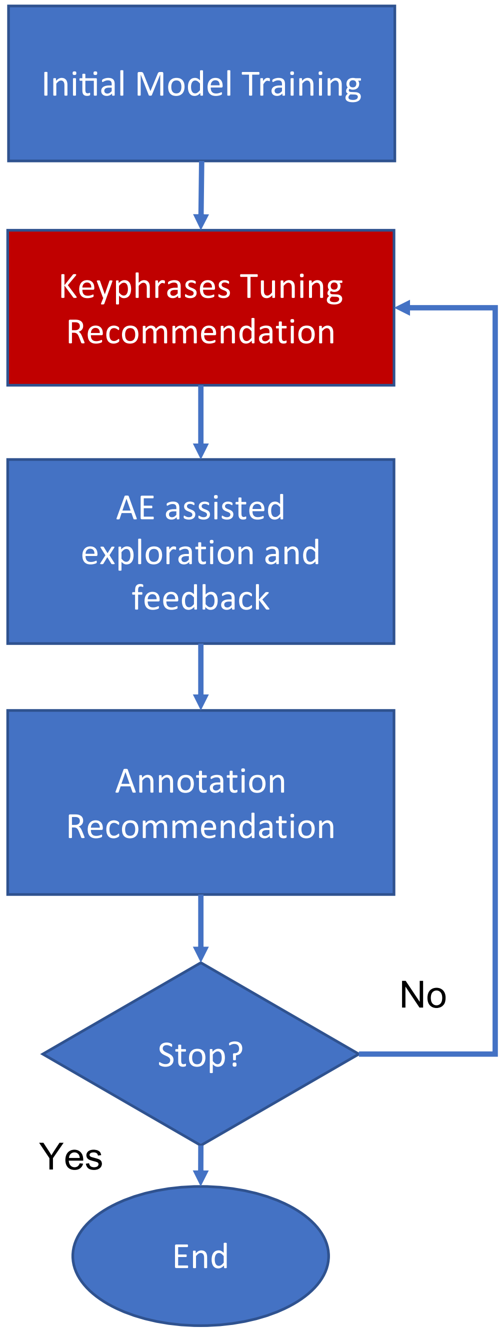


Figure 2. Automatic Exploration Process

Component 2.3: AE assisted exploration and feedback

Figure 3. Interactive Auto Exploration Interface (IAEI)

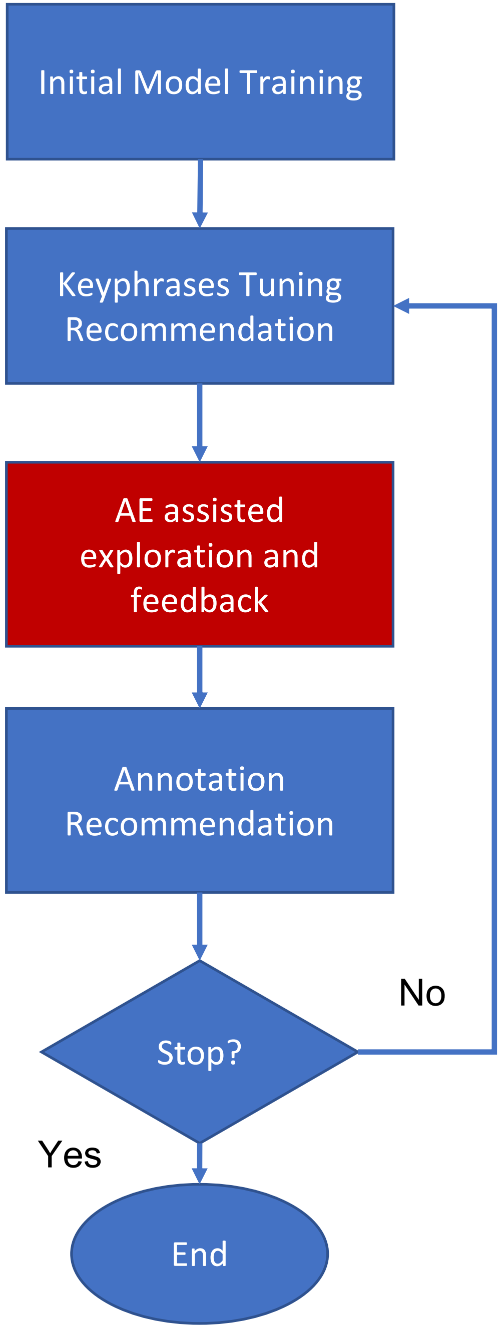


Figure 2. Automatic Exploration Process

Component 2.4: Annotation and stop recommendations

- Annotation recommendation
  - We rank the unannotated posts according to their Aggregated MI(AMI) score
  - Pick up top n(10 in our case) to let annotator do more annotation
- Stop recommendation
  - Accuracy of ML not increase
  - normalized discounted cumulative gain (nDCG) between the keyword's MI score ranking (descending) and the keyword I score ranking not increase.

$$AMI(p) = \sum_{i=1}^{n} MI(t_i, c_t)$$

Figure 5. Aggregated MI score formula

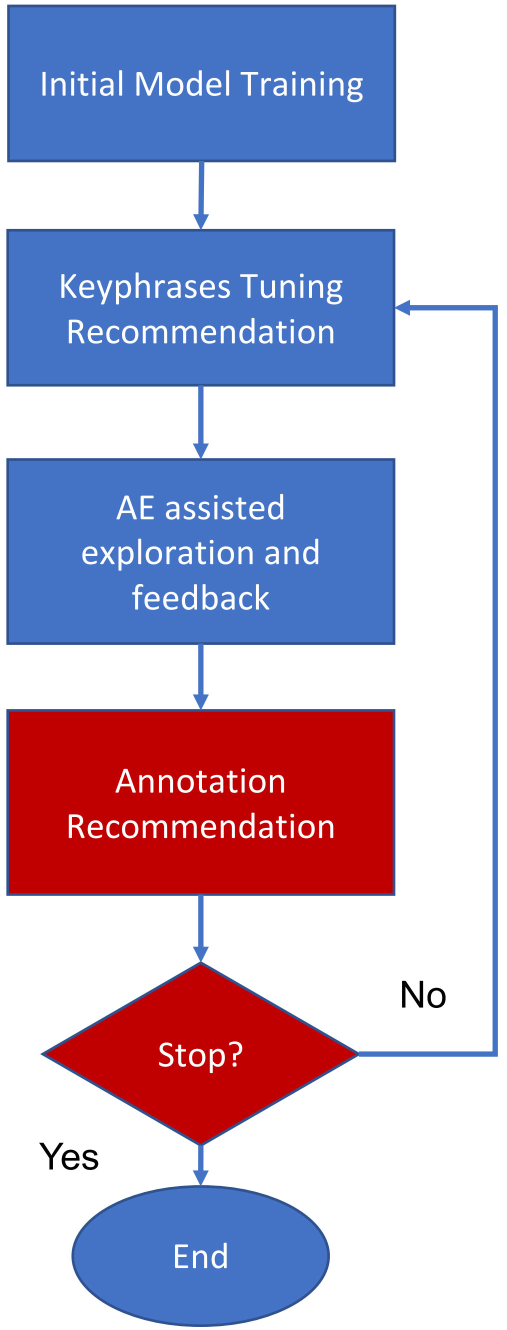


Figure 2. Automatic Exploration Process

Experiment Results

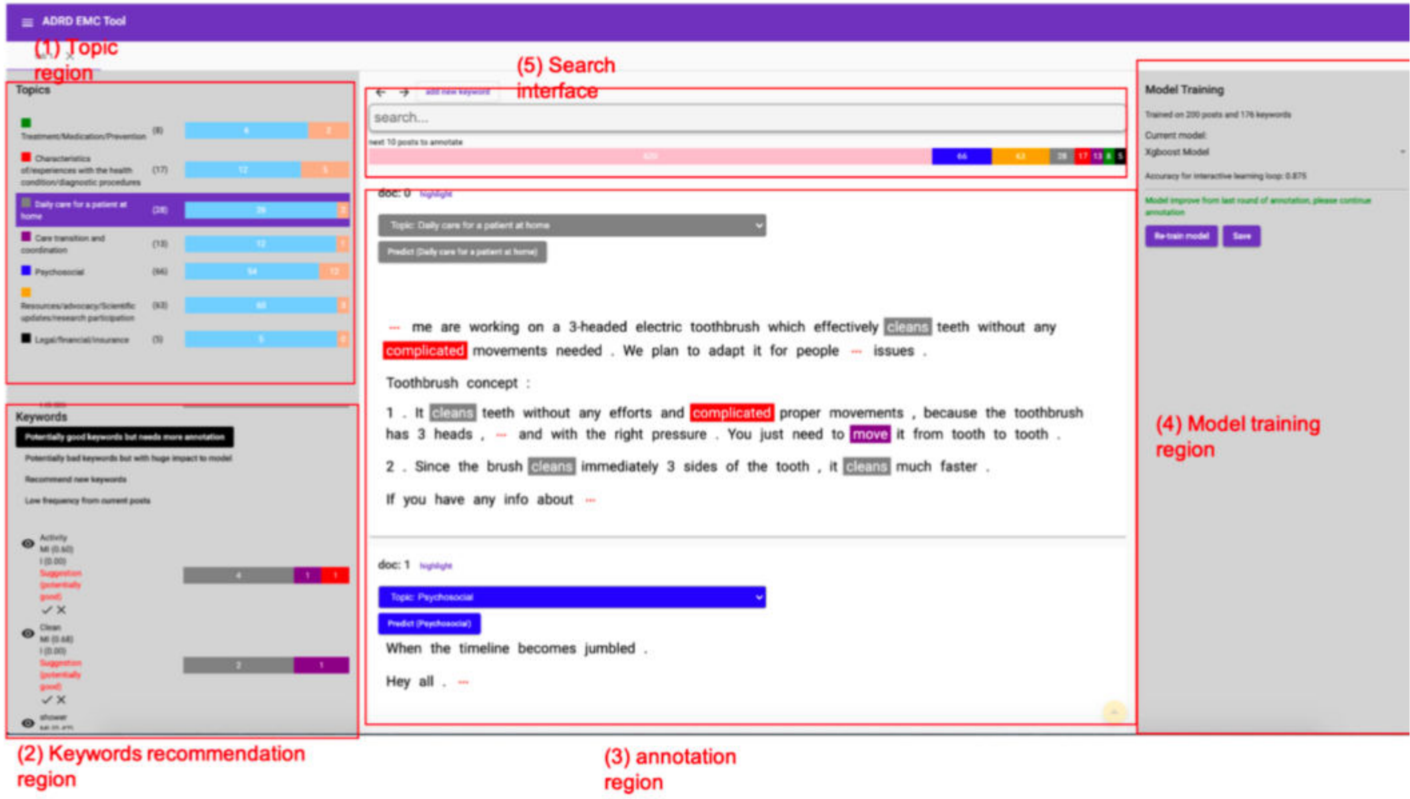
- Result
  - IML improve the performance
  - KT doesn't always improve the performance but with IML, it achieves best performance
  - Human experts annotated just 40 more posts
- HIW-ADRD 3.1 framework
  - discovered 7 better keyphrases to replace the existing one
  - removed 6 keyphrases
  - reviewed 25 new keyphrases and include 15 of them

Training group	Dataset	Xgboost	SVM	Naïve Bayes
ID	TRAIN	0.870	0.640	0.705
	TEST	0.281	0.351	0.281
ID+RD (benchmark)	TRAIN	0.854	0.618	0.699
	TEST	0.333	0.404	0.333
ID + IML	TRAIN	0.887	0.637	0.692
	TEST	0.509	0.421	0.386
ID+RD+KT (benchmark)	TRAIN	0.851	0.630	0.696
	TEST	0.316	0.439	0.316
ID+IML+KT	TRAIN	0.892	0.658	0.696
	TEST	**0.544***	**0.456**	**0.421**

Table 3. model accuracy

Usability Interview with Human Experts

- Strengths
  - the visualization of keyphrase recommendations and navigation
  - The category distribution for search
  - The highlight keyphrases in posts
- Weaknesses
  - It takes time for domain experts to understand MI and I score
  - Some keyphrases recommendation is bad because of word normalization



Takeaway messages:

- Interactive ML can be used to enable expert machine collaboration
- Keyphrases can be the focus of the interactions to enable the collaboration

Outline

- Basics of Keyphrases: Definitions and Importance
- Identification of Keyphrases: Extraction and Generation
- Applications of Keyphrases: knowledge unit for supporting student learning
- Applications of Keyphrases: knowledge unit for recognizing patients’ concerns
- Applications of Keyphrases: knowledge unit for interactive machine learning
- Conclusions

Conclusions

- ▷ Keyphrases are short noun phrases to summarize and highlight important information in a piece of text
  - Different to and related to words and concepts
- ▷ Keyphrases can provide unique contributions as computable knowledge unit
  - Only a few possible applications are presented, many more can be explored
- ▷ Keyphrases can take in different roles in the text
  - Still open questions on how to identify and make use of their roles

Acknowledgement

Collaborators who made the mentioned work possible

▷ School of Computing and Information

- Rui Meng, Khushboo Thaker, Zhimeng Luo, Hung Chau, Zhendong Wang, Ning Zou, Mengdi Wang, Peter Brusilovsky, Diyang Xue

▷ UPMC

- Adam Frisch

▷ University of Texas Austin

- Bo Xie, Robin C. Hilsabeck, Alyssa Aguirre

Funding Agencies that partially supported the work

- National Science Foundation, Amazon, National Institute of Health, University of Pittsburgh

Thank you & Questions