Deep Keyphrase Generation

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, Yu Chi
School of Computing and Information
University of Pittsburgh

Keyphrase

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-

- **Keyphrase**
  - Short texts highly summarize the significant content of a document
  - Applications
    - Knowledge mining (concept)
    - Information retrieval (indexing term)
    - Summarization
  - Provided by authors/editors

- **This work aims to**
  - obtain **keyphrases** from scientific papers (title+abstract) automatically

Previous Approaches

- **3-step process**

Source Text

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-

5. negative impact (0.057)

Issues:

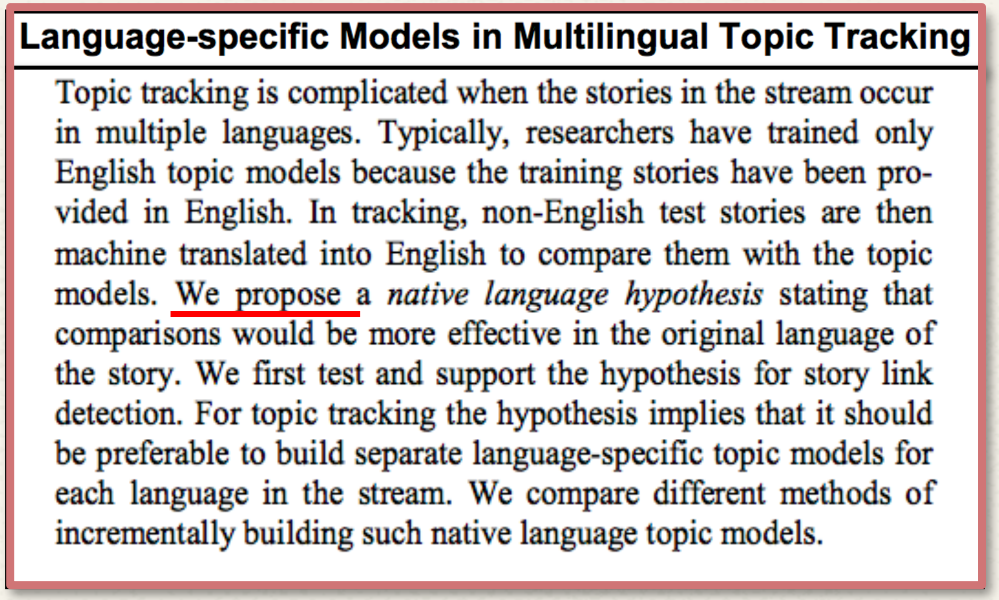
1. Candidates must be acquired from the source text.
   - Only able to predict phrases appear in text
   - **Present** & **Absent**
2. Highly rely on manual feature design
   - simple features can hardly represent deep semantics
   - neither flexible nor scalable

Dataset	% Present	% Absent
Inspec	73.62%	26.38%
Krapivin	54.33%	45.67%
NUS	45.63%	54.37%
SemEval	55.66%	44.34%

Performance Upper Bound

Revisit Keyphrase Generation

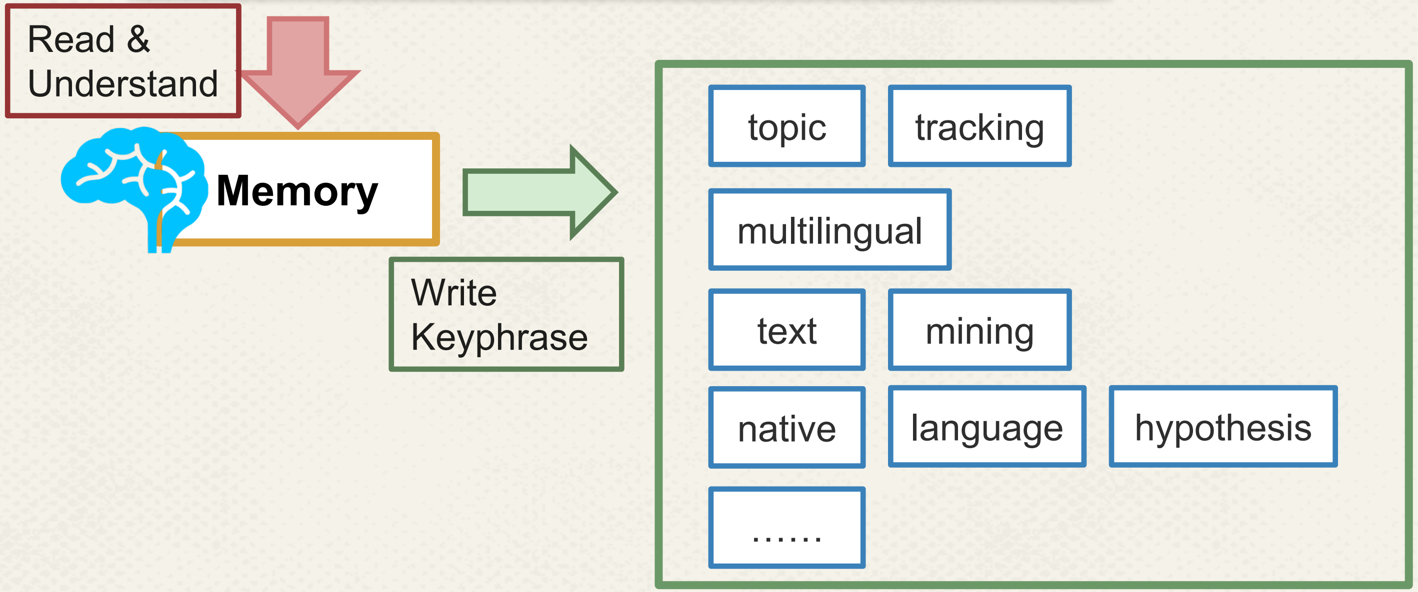
- **How do humans assign keyphrases?**
  1. Reading the text
  2. Understand and get contextual information
  3. Summarize and write down the most meaningful phrases
  4. Get hints from text, copy certain phrases



Language-specific Models in Multilingual Topic Tracking

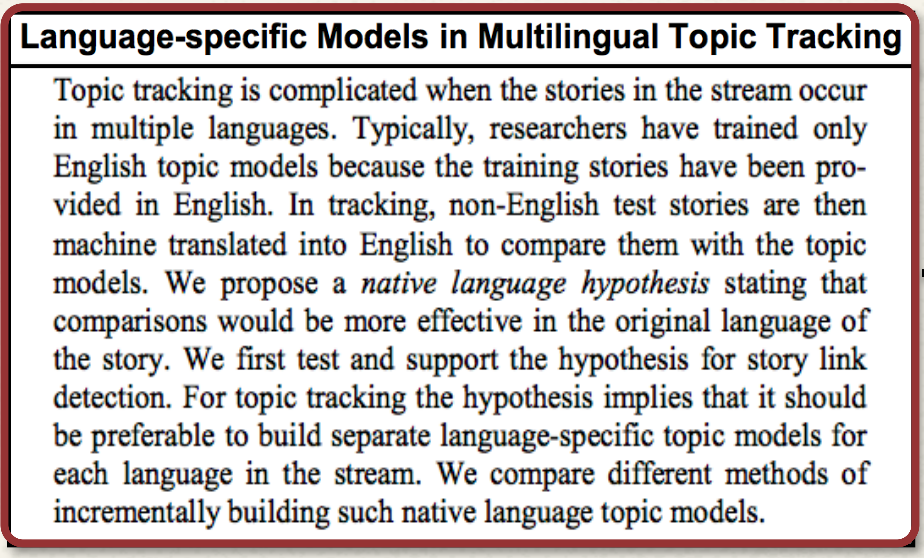
Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare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

- **Can machine simulate this process?**
  - Recurrent Neural Networks [Step 1-3]
  - Copy Mechanism [Step 4]



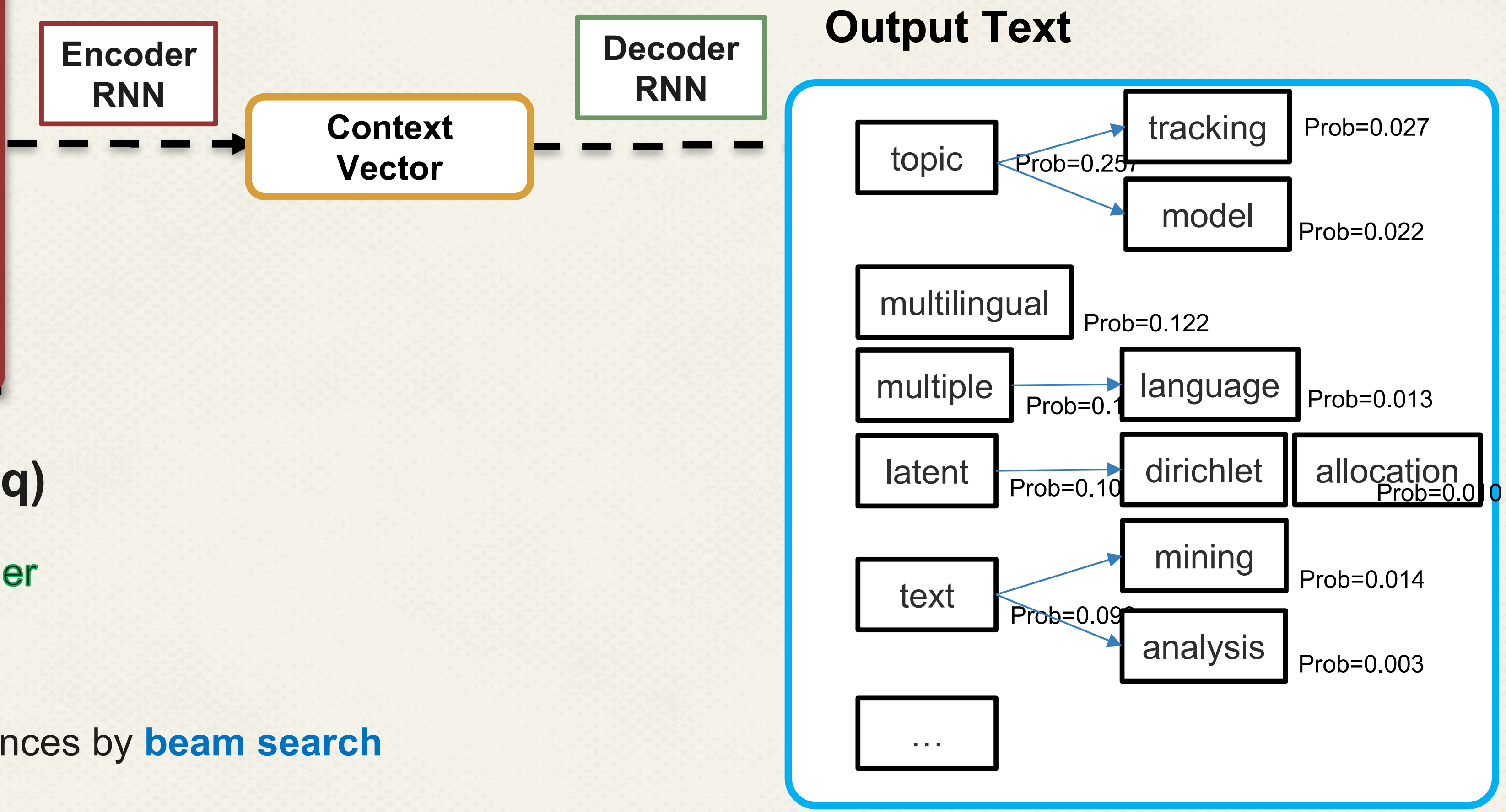
Recurrent Neural Networks

Input Text



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. 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.

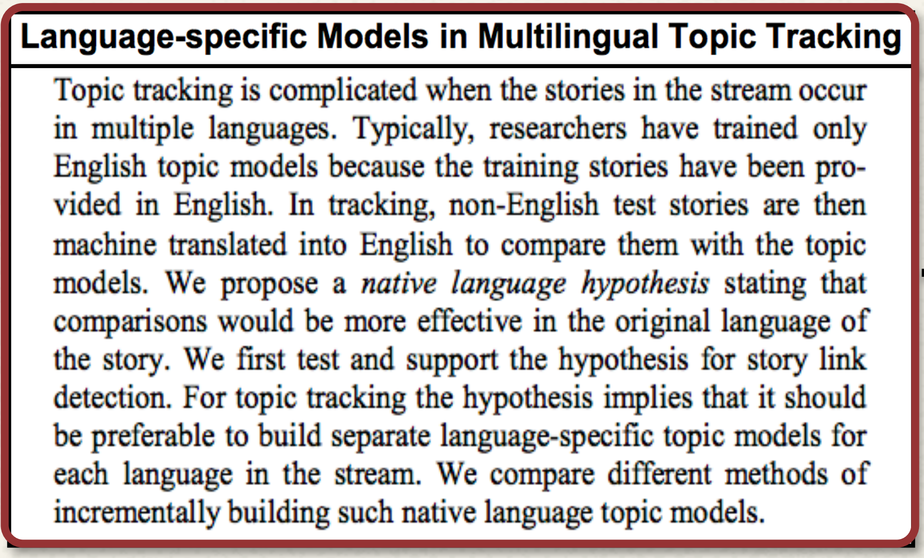


Encoder-decoder model (Seq2seq)

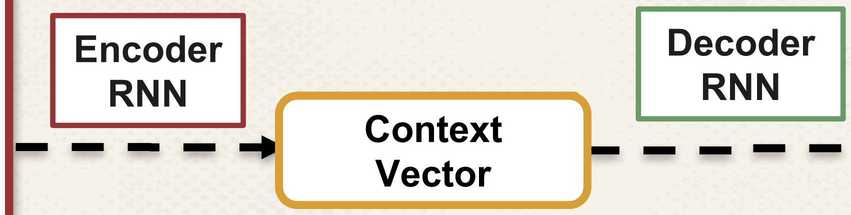
- **One RNN encoder and one RNN decoder**
- Gated recurrent units (GRU) cell
- Decoder generates multiple short sequences by **beam search**

Recurrent Neural Networks

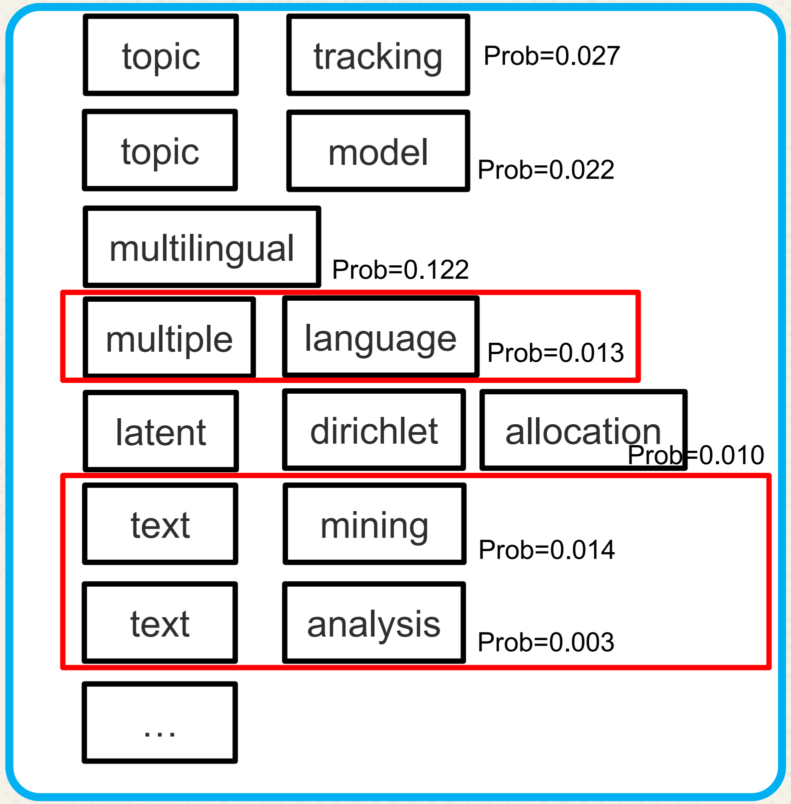
Input Text


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. 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.



Output Text

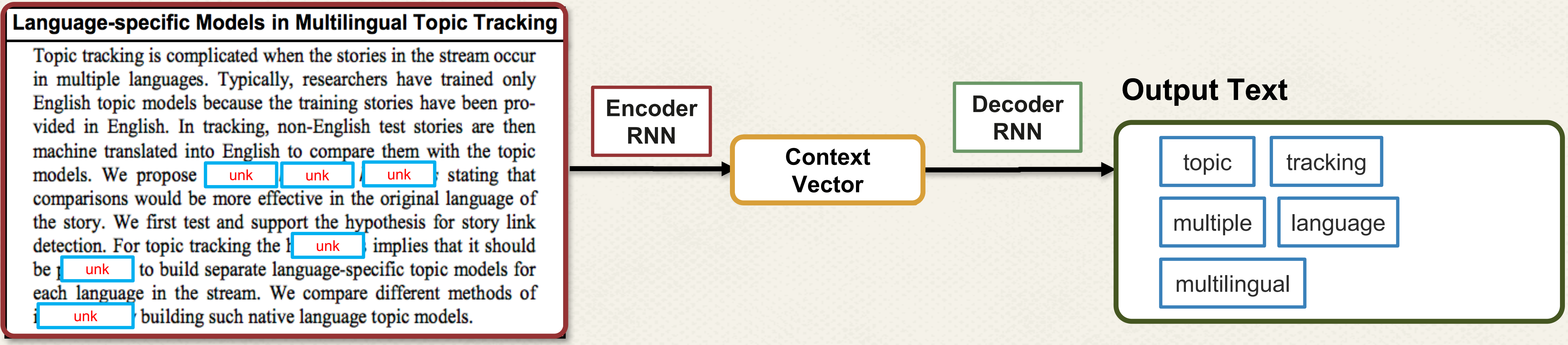


Encoder-decoder model (Seq2seq)

- **One RNN encoder and one RNN decoder**
- Gated recurrent units (GRU) cell
- Decoder generates multiple short sequences by **beam search**
- Rank them and return the **top K results**

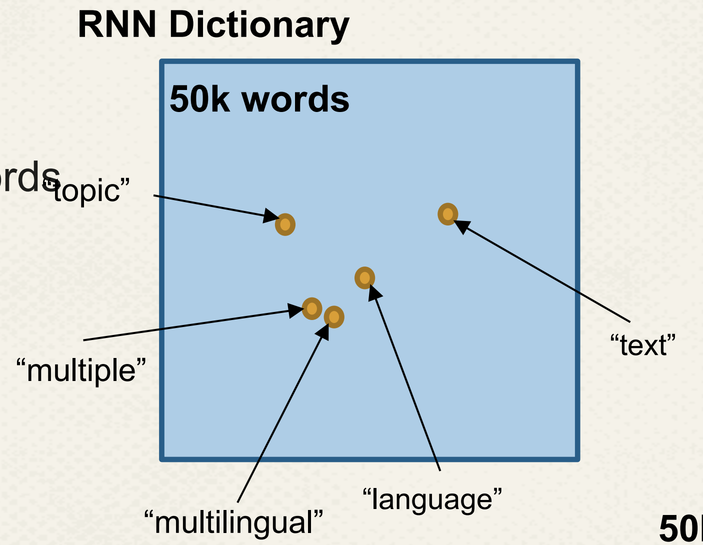
Recurrent Neural Networks

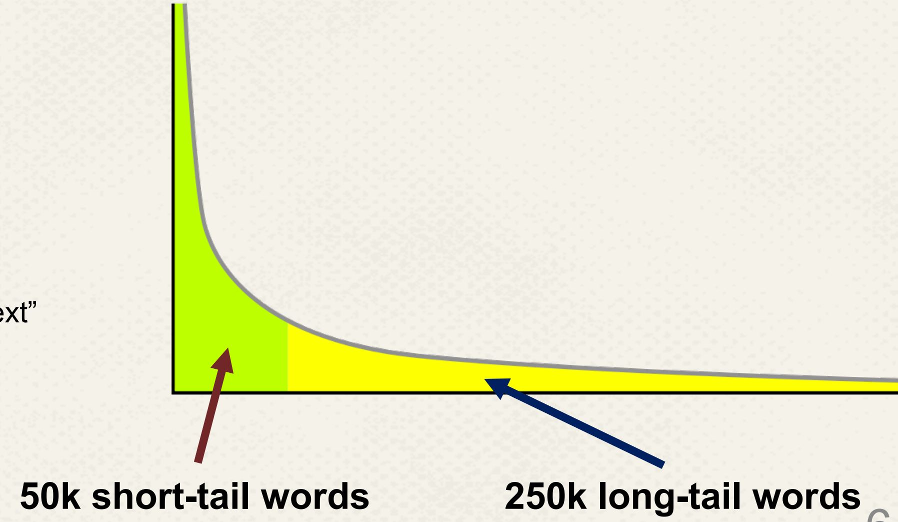
Input Text



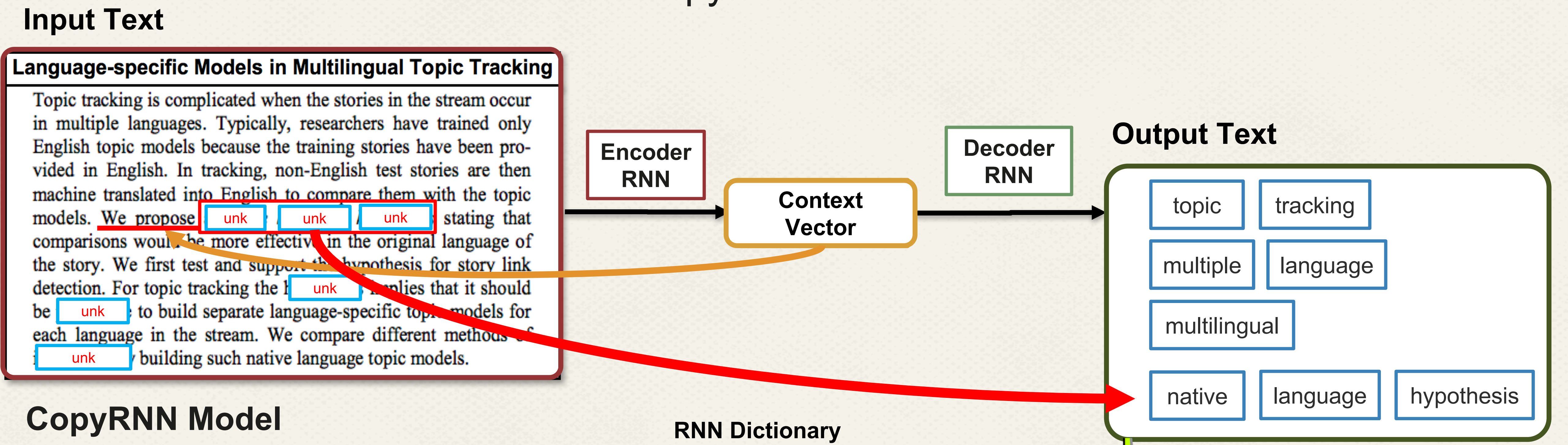
Problem of RNN model

- Keep everything in memory
- Only train vectors for top 50k high-frequency words
- Long-tail words are replaced with an “unknown” symbol <unk>
    - Unable to predict long-tail words
    - Many keyphrases contain long-tail words (2%)



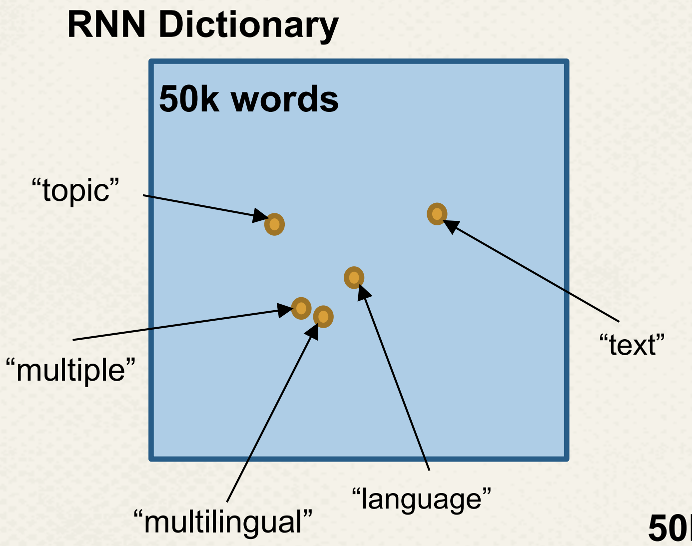


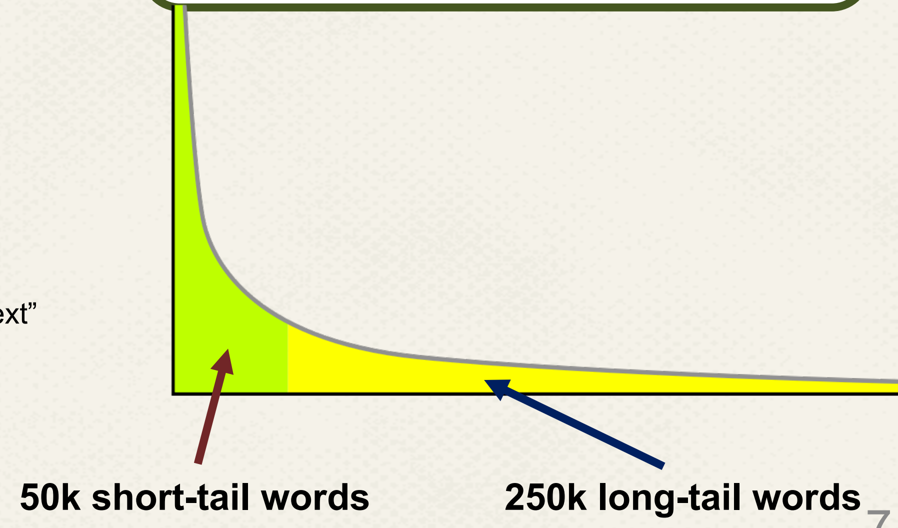
Copy Mechanism



CopyRNN Model

- Copy words from input text
- Locate the words of interest by <u>contextual</u> features
- <u>Copy</u> corresponding part to output
- Enhance the RNN with <u>extractive ability</u>





Dataset

- All data are scientific papers in Computer Science domain
- **Training Data**
  - Collected from Elsevier, ACM Digital Library, Web of Science etc.
    - # (Paper) = 571,267
    - # (Phrase) = 3,011,651
    - # (Unique word) = 324,163
- **Testing Data**
  - Four commonly used datasets, only use abstract text
  - Overlapping papers are removed from training dataset

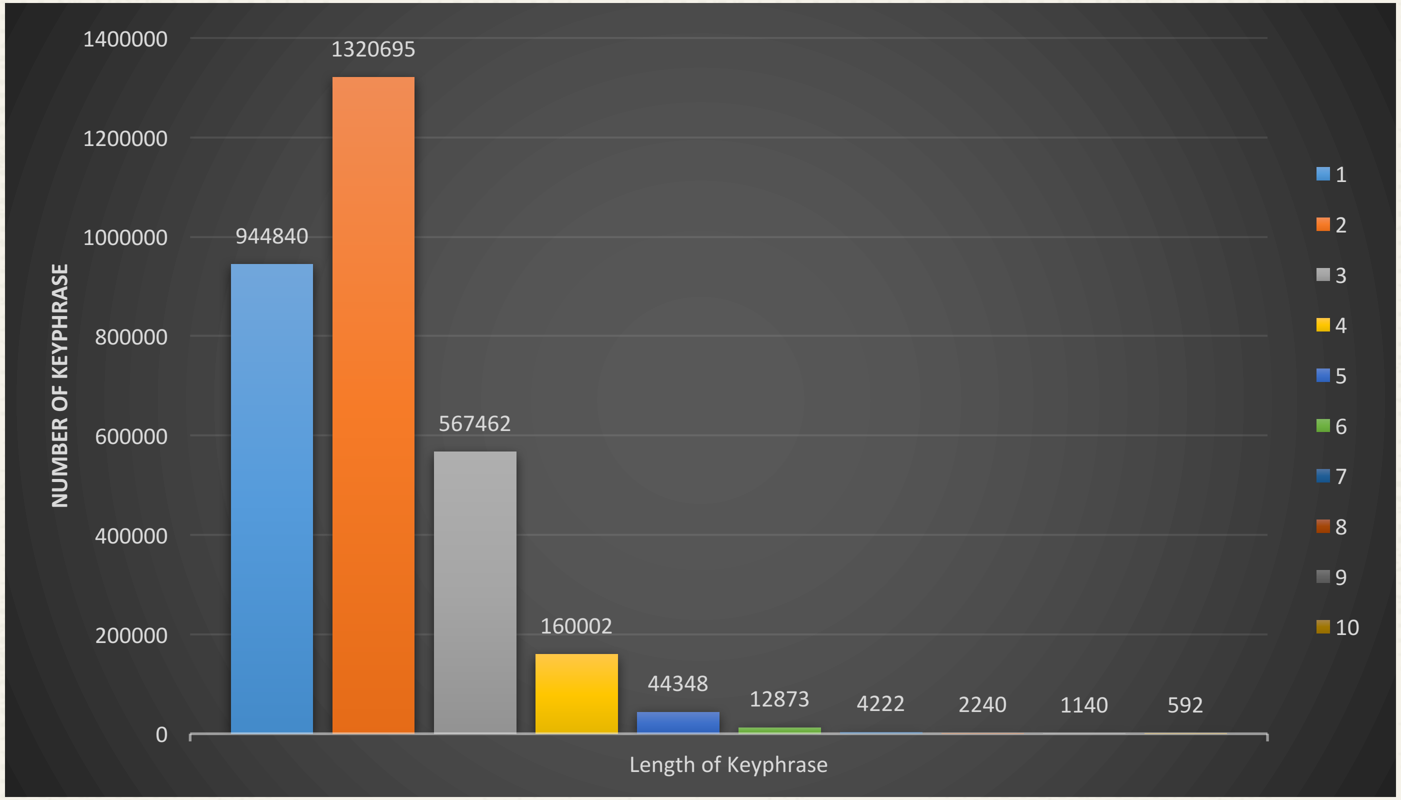
Dataset	# Paper	# All (Avg)	# Present	# Absent	% Absent
Inspec	500	4,913 (9.82)	3,617	1,296	26.38%
Krapivin	400	2,461 (6.15)	1,337	1,124	45.67%
NUS	211	1,466 (6.94)	669	797	54.37%
SemEval	100	2,339 (23.39)	1,302	1,037	44.34%
KP20k	20,000	105,471 (5.27)	66,221	39,250	37.21%

Experiment

Dataset

#(Unique Keyphrase)=324,163

Length of Terms	Number of Frequency	Percentag e
1	944840	30.88%
2	944840	43.16%
3	567462	18.55%
4	160002	5.23%
5	44348	1.45%
>5		0.73%



Experiment Setup

- **Evaluation Methods**
  - Process ground-truth and predicted phrases with Porter stemmer
  - Macro-average of precision, recall and F-measure @5,@10
- **Tasks**

1. **Present phrases** prediction
   - Compare to previous studies: Tf-Idf, TextRank, SingleRank, ExpandRank, KEA, Maui
2. **Absent phrases** prediction
   - No baseline comparison
3. Transfer to news dataset

Result

Task 1 - Predict Present Keyphrase

Dataset	Inspec		Krapivin		NUS		SemEval		KP20k	
Method	F@5	F@10	F@5	F@10	F@5	F@10	F@5	F@10	F@5	F@10
Tf-Idf	0.221	0.313	0.129	0.160	0.136	0.184	0.128	0.194	0.102	0.126
TextRank	0.223	0.281	0.189	0.162	0.195	0.196	0.176	0.187	0.175	0.147
SingleRank	0.214	0.306	0.110	0.153	0.140	0.173	0.135	0.176	0.096	0.119
ExpandRank	0.210	0.304	0.110	0.152	0.132	0.164	0.139	0.170	-	-
KEA	0.098	0.126	0.123	0.134	0.069	0.084	0.025	0.026	0.171	0.154
Maui	0.040	0.042	0.249	0.216	0.249	0.268	0.044	0.039	0.270	0.230
RNN	0.085	0.064	0.135	0.088	0.169	0.127	0.157	0.124	0.179	0.189
CopyRNN	**0.278** (24.7%)	**0.342** (9.3%)	**0.311** (24.9%)	**0.266** (23.1%)	**0.334** (34.1%)	**0.326** (21.6%)	**0.293** (66.5%)	**0.304** (56.7%)	**0.333** (23.3%)	**0.262** (13.9%)

Take-away

1. Naïve RNN model fails to compete with baseline models
2. CopyRNN models outperform baseline models and RNN significantly. Copy mechanism can capture key information in source text.

Example - Phraseness

[Title]
Nonlinear Extrapolation Algorithm for Realization of a Scalar Random Process
[Abstract]
A method of construction of a nonlinear extrapolation algorithm is proposed. This method makes it possible to take into account any nonlinear random dependences that exist in an investigated process and are described by mixed central moment functions. The method is based on the V. S. Pugachev canonical decomposition apparatus. As an example, the problem of nonlinear extrapolation is solved for a moment function of third order.
[Ground-truth] 6 ground-truth phrases

moment function	nonlinear extrapolation algorithm	canonical decomposition apparatus
scalar random process	nonlinear random dependences	mixed central moment functions

[Prediction]

Tf-Idf

account
example
method
mixed central moment functions
<u>**moment function**</u>
nonlinear extrapolation
<u>**nonlinear extrapolation algorithm**</u>
<u>**nonlinear random dependences**</u>
problem
process
pugachev canonical decomposition apparatus
realization
s
scalar random process
third order

CopyRNN

nonlinear extrapol
<u>**moment function**</u>
canon decomposit
extrapol algorithm
<u>**scalar random process**</u>
random process
central moment function
<u>**nonlinear extrapol algorithm**</u>
<u>**mix central moment function**</u>
central moment
mix central moment
random depend
investig process
<u>**nonlinear random depend**</u>
scalar random

Result

Example – Failure of RNN

[Title]
Meta-level Coordination for Solving Distributed Negotiation Chains in Semi-cooperative Multi-agent Systems
[Abstract]
A negotiation chain is formed when multiple related negotiations are spread over multiple agents. In order to appropriately order and structure the negotiations occurring in the chain so as to optimize the expected utility, we present an extension to a single-agent concurrent negotiation framework. This work is aimed at semi-cooperative multi-agent systems, where each agent has its own goals and works to maximize its local utility; however, the performance of each individual agent is tightly related to other agents' cooperation and the system's overall performance. We introduce a pre-negotiation phase that allows agents to transfer meta-level information. Using this information, the agent can improve the accuracy of its local model about how other agents would react to the negotiations …
[Ground-truth] 7 ground-truth phrases
multipl agent; negoti framework; negoti chain; semi cooper multi agent system; pre negoti; agent; flexibl;
[Prediction]

Tf-Idf	RNN	CopyRNN
pre negoti phase	multi agent system	multi agent system
semi cooper multi agent system	multi agent	**negoti chain**
system s overal perform	multiag system	multiag system
negoti	agent system	concurr negoti
negoti chain	**multipl agent**	artifici intellig
individu agent	artifici intellig	**pre negoti**
other agent s cooper	cooper multi agent system	multi agent
concurr negoti framework	cooper multi agent	**semi cooper multi agent system**
cooper multi agent system		**multipl agent**
multipl relat negoti		expect util
negoti chain		distribut artifici intellig
meta level coordin		global negoti
negoti solut		meta level coordin
global negoti chain context		semi cooper

Example – Phrases with OOD words

[Title]
Full-screen ultrafast video modes over-clocked by simple VESA routines and registers reprogramming under MS-DOS.

[Abstract]
Fast full-screen presentation of stimuli is necessary in psychological research. Although Spitczok von Brisinski (1994) introduced a method that achieved ultrafast display by reprogramming the registers, he could not produce an acceptable full-screen display. In this report, the author introduces a new method combining VESA routine calling with registers reprogramming that can yield a display at 640 × 480 resolution, with a refresh rate of about 150 Hz.

[GROUND-TRUTH] 6 ground-truth phrases
vesa routine calling; fast full screen stimuli presentation; ms dos; full screen ultrafast video modes; psychological research ; register reprogramming;

[**PREDICTION**]

1. **register reprogramming**	2. video modes
3. ultrafast display	4. screen display
5. ultrafast video	6. **vesa** routine [copied]
7. refresh rate	8. routine calling
9. ultrafast video modes	10. **psychological research**
11. vesa routine calling [copied]	12. **spitczok** von[copied]
13. video modes over clocked	14. **spitczok** von brisinski[copied]

- Nearly 2% of all the correct predictions contain out-of-vocabulary words

Task 2 - Predict Absent Keyphrase

- Same five test datasets, only use absent keyphrases as ground-truth
- Evaluate with recall@10 and recall@50

Dataset	RNN		CopyRNN+	
	Recall @10	Recall @50	Recall @10	Recall @50
Inspec	0.0309	0.0610	**0.0471**	**0.0995**
Krapivin	0.0945	0.1562	**0.1128**	**0.2015**
NUS	0.0498	0.0890	**0.0578**	**0.1157**
SemEval	0.0414	0.0602	**0.0427**	**0.0665**
KP20k	0.0833	0.1441	**0.1253**	**0.2108**

Result

Task 2 - Predict Absent Keyphrase

[Title]

Towards **content-based** relevance **ranking** for video search

[Abstract]

Most existing web **video** search engines **index videos** by file names, URLs, and surrounding texts. These types of **video** metadata roughly describe the whole video in an abstract level without taking the rich content, such as **semantic** content descriptions and **speech** within the video, into consideration. Therefore the relevance ranking of the video search results is not satisfactory as the details of video contents are ignored. In this paper we propose a novel relevance ranking approach for Web-based video search using both video metadata and the rich content contained in the videos. To leverage real content into ranking, the **videos are segmented** into shots, which are smaller and more semantic-meaningful retrievable units, and then more detailed information of video content such as semantic descriptions and **speech** of each shots are used to improve the retrieval and ranking performance. With video metadata and content information of shots, we developed an integrated ranking approach, which achieves improved ranking performance. We also introduce machine learning into the ranking system, and compare them with IR-model (information retrieval model) based method. The evaluation results demonstrate the effectiveness of the proposed ranking methods.

[Ground-truth] 10 absent phrases

video segmentation, ir model, content based approach, content based ranking, neutral network based ranking, video index, learning based ranking, ir model based ranking, machine learning model, video retrieval

[Predictions]

1. video retrieval [correct!] 2. web search 3. content ranking 4. content based retrieval 5. content retrieval
6. video indexing [correct!] 7. relevance feedback. 8. video ranking 9.semantic web 10. content based video retrieval
11. web metadata 12. video analysis 13. speech recognition 14. content analysis 15. speech retrieval

34. content based ranking [correct!] 61. video segmentation [correct!]

Task 3 – Transfer to News Articles

- So far training and testing are only about scientific papers
- What if transfer it to a completely unseen domain
    - Does model learn any universal feature?

- Test the CopyRNN on DUC-2001
    - 308 news articles and 2,488 keyphrases
    - CopyRNN recalls 766 keyphrases. 14.3% contain out-of-vocabulary words
    - Many names of persons and places are correctly predicted.

Model	F1-score
TFIdf	0.270
TextRank	0.097
SingleRank	0.256
ExpandRank	0.269
KeyCluster	0.140
CopyRNN@10	0.164

Example – Transfer to News Articles

[Article]
anti maoists threaten prosecutor. a **death squad** opposed to the shining path guerrillas has threatened to kill a district attorney if he investigates charges that soldiers massacred dozens of peasants , his office said tuesday . police said members of shining path , a maoist group , killed two policemen and wounded three in jungle raids . the rodrigo franco command , which has vowed to kill a shining path member or sympathizer for every person slain by guerrillas , issued the threat against district attorney carlos escobar on monday , according to his office in andean city of ayacucho . escobar is investigating charges that troops rounded up dozens of peasants , accused them of being shining path members and killed them . the alleged massacre occurred in may near cayara , a farming village <digit> miles south of ayacucho . officials said the **rebel raids** occurred sunday , at a police post and telephone relay station near the jungle city of pucallpa , <digit> miles northeast of lima . shining path guerrillas began fighting eight years ago . the government says more than <digit> , <digit> people have been killed and puts the property damage at <digit> billion . the rodrigo franco group is named for an official of the government party killed the shining path killed last year . it became known in july when it claimed responsibility for killing the lawyer for **osman morote** . he is suspected of being the shining path second in command and is in jail on terrorism charges .

[Ground-truth] 8 present phrases
shining path guerrillas; police post; rebel raids; death squad; property damage; rodrigo franco command; district attorney carlos escobar; osman morote;

[Predictions]

1. shining path
2. **death squad**[correct]
3. district attorney
4. **rebel raids**[correct]
5. **osman morote**[correct]
6. jungle raids
7. rodrigo franco
8. terrorism charges
9. relay station
10. anti maoists
11. massacred dozens
12. andean city

Conclusion & Future Work

- Keyphrase generation study based on deep learning methods
  - First work concerns absent keyphrase prediction
  - RNN + Copy mechanism
  - Able to learn cross-domain features

- Better model on capturing contextual information
- Multiple-output optimization
- Long documents, length & diversity penalties on output sequences

THANKS!

Any question?