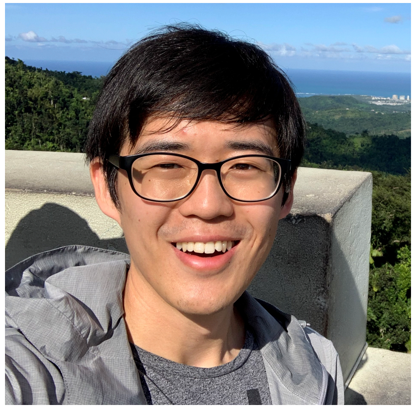
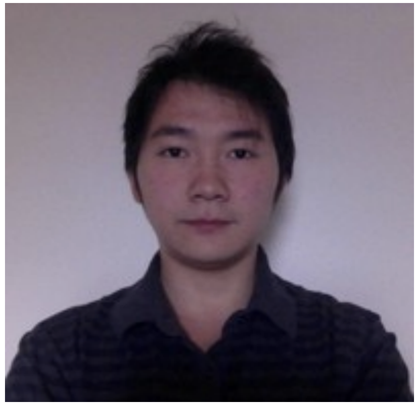
An Empirical Study on Neural Keyphrase Generation



Rui Meng

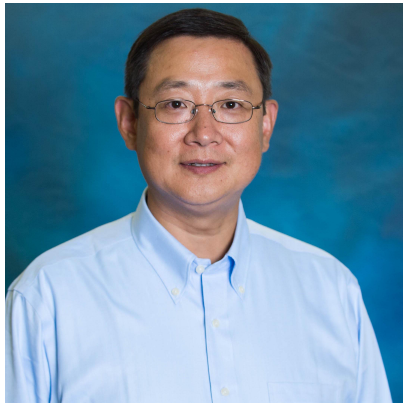
Xingdi Yuan

Tong Wang



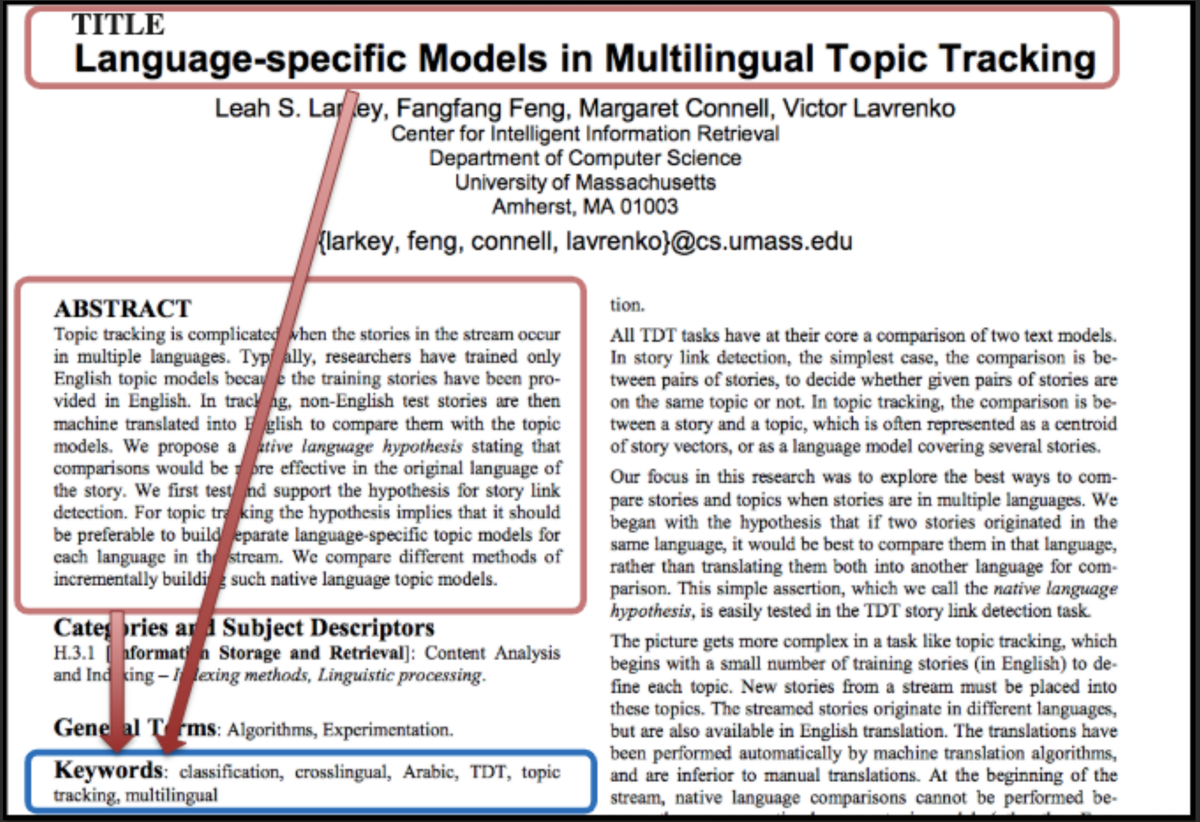
Sanqiang Zhao

Adam Trischler



Daqing He

What’s KPG & why it’s unique?


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.

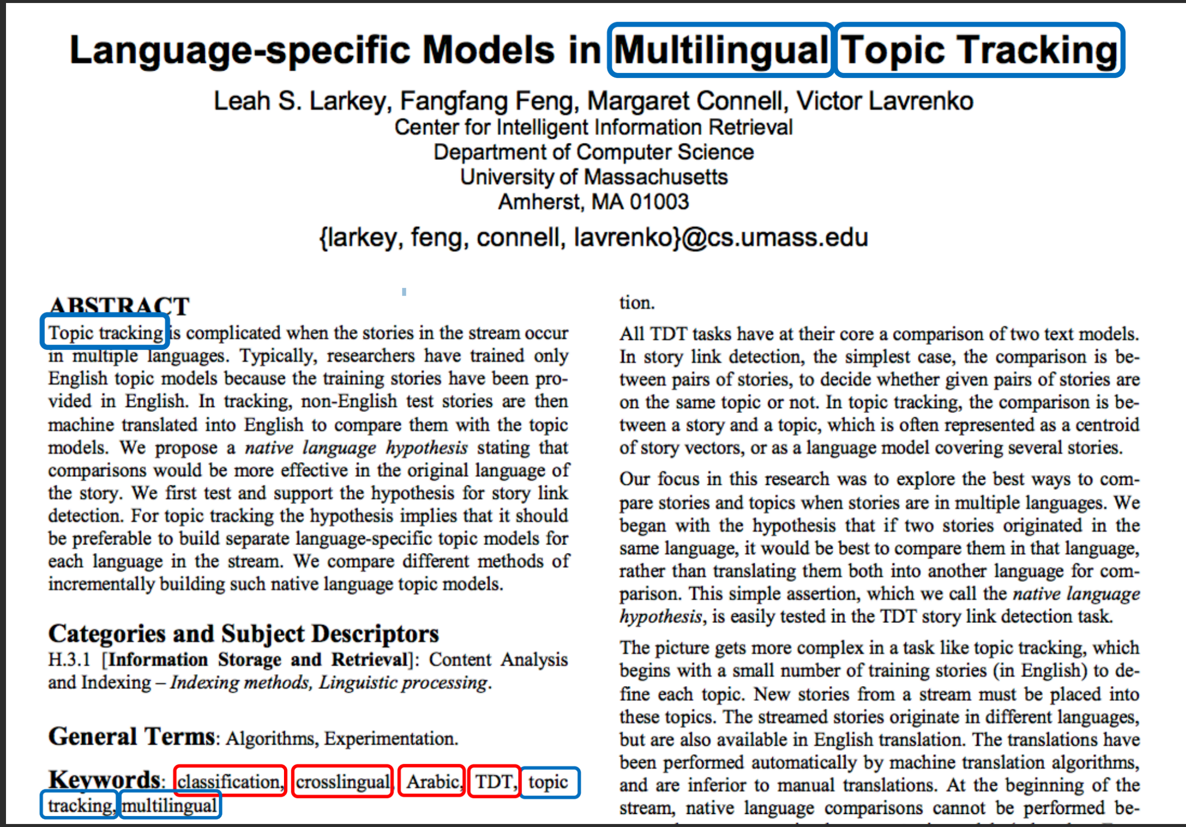
All TDT tasks have at their core a comparison of two text models. In story link detection, the simplest case, the comparison is between pairs of stories, to decide whether given pairs of stories are on the same topic or not. In topic tracking, the comparison is between a story and a topic, which is often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

- Important concepts/entities in a document.
- Each phrase can have multiple words
- Target is a list of multiple phrases (variable number of target sequences)

What’s KPG & why it’s unique?



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-

- Present (extractive) vs Absent (abstractive)

Seq2Seq for KPG: One2One vs One2Seq

- One2One:

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

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

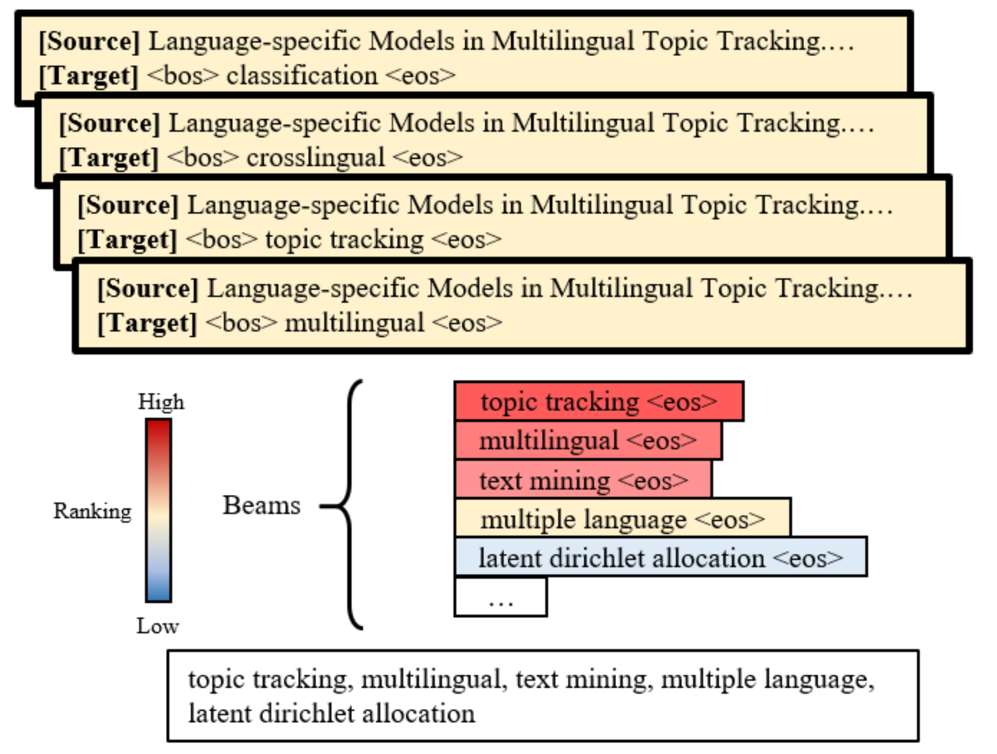
[Source] Language-specific Models in Multilingual Topic Tracking….
[Target] <bos> topic tracking <eos>

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

- Deep keyphrase generation. R Meng, S Zhao, S Han, D He, P Brusilovsky, Y Chi, 2017.

Seq2Seq for KPG: One2One vs One2Seq

- One2One:

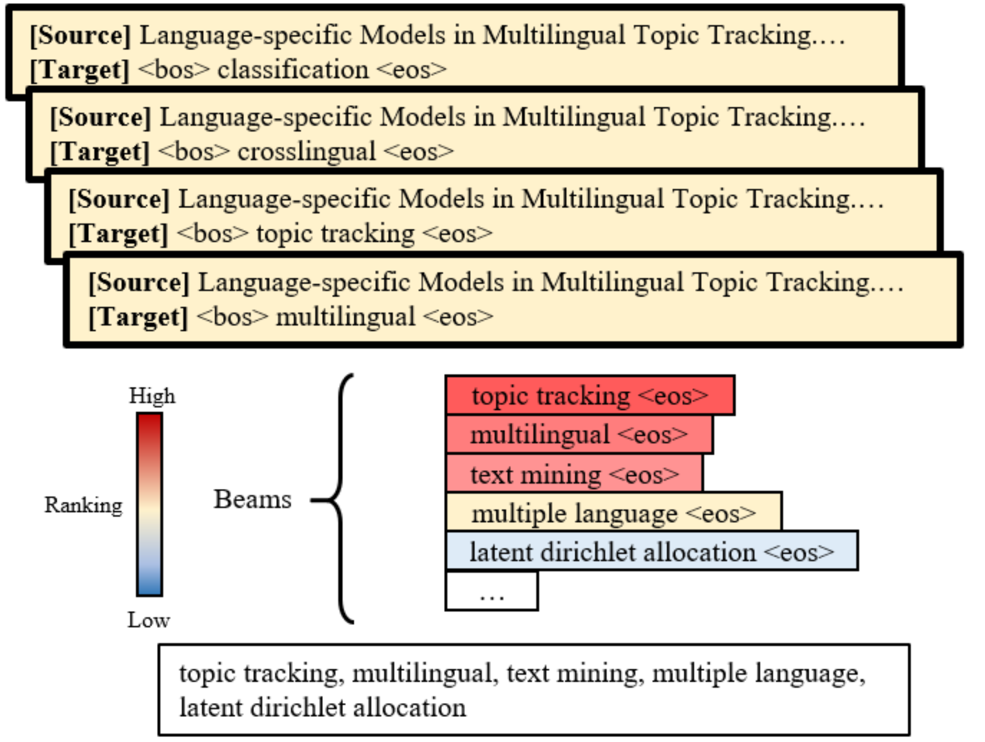


(top 5) beam search outputs

- Deep keyphrase generation. R Meng, S Zhao, S Han, D He, P Brusilovsky, Y Chi, 2017.

Seq2Seq for KPG: One2One vs One2Seq

- One2One:



(top 5) beam search outputs

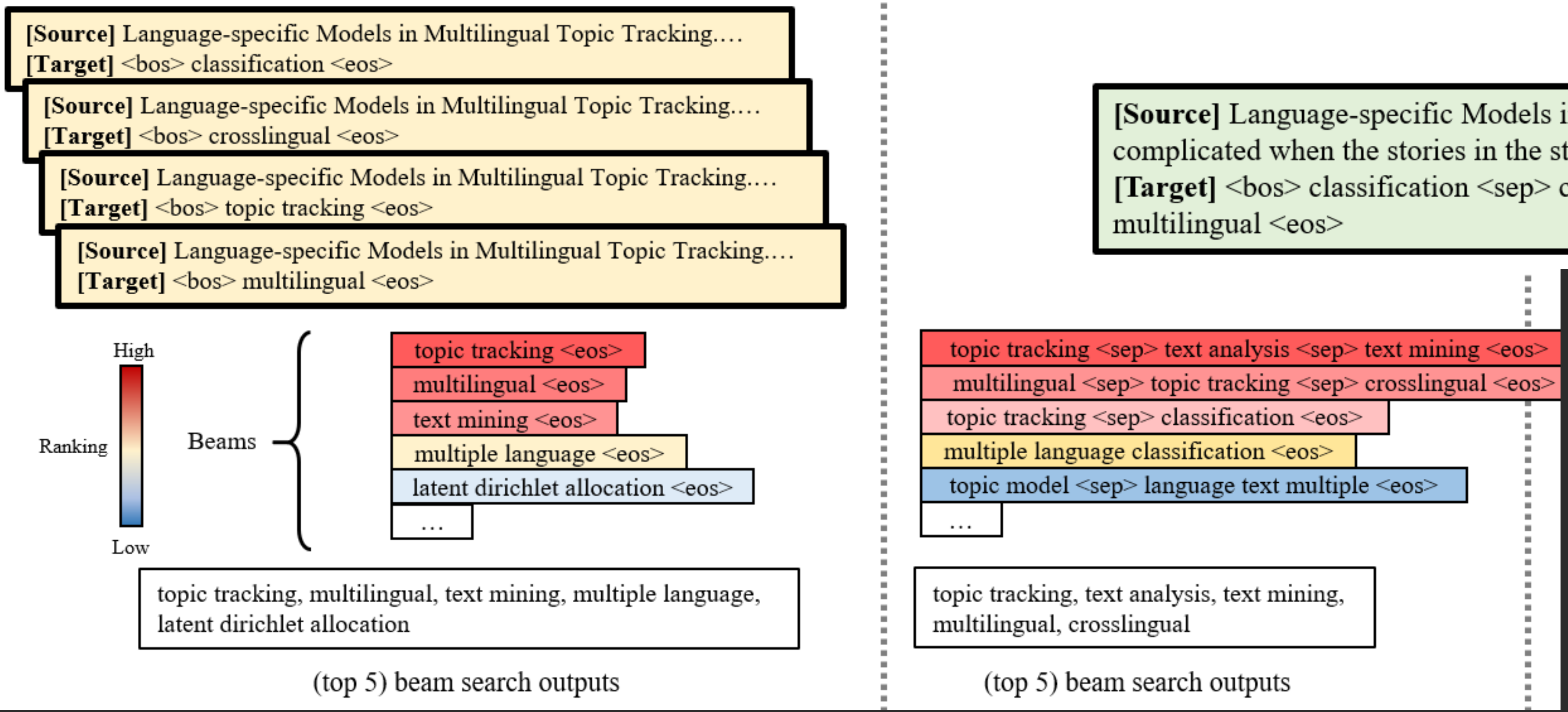
- One2Seq:

[Source] Language-specific Models in Multilingual Topic Tracking. Topic tracking is complicated when the stories in the stream occur in multiple languages....
[Target] <bos> classification <sep> crosslingual <sep> topic tracking <sep> multilingual <eos>

- Deep keyphrase generation. R Meng, S Zhao, S Han, D He, P Brusilovsky, Y Chi, 2017.
- One size does not fit all: Generating and evaluating variable number of keyphrases. X Yuan, T Wang, R Meng, K Thaker, P Brusilovsky, D He, A Trischler, 2020.

Seq2Seq for KPG: One2One vs One2Seq

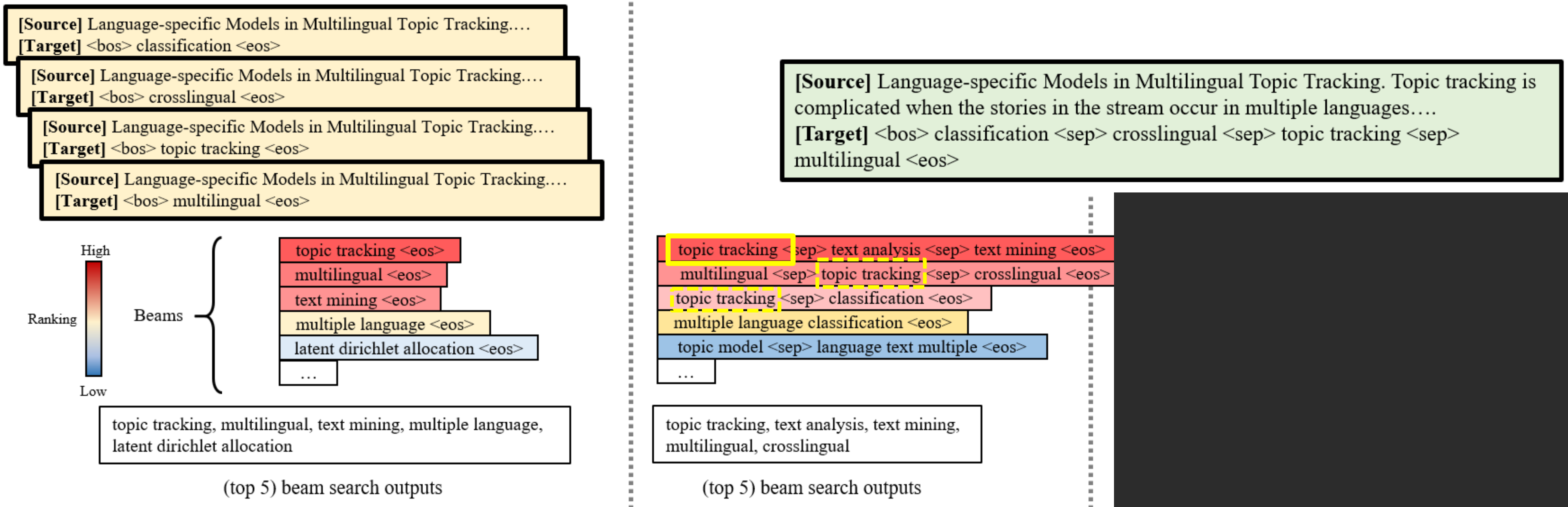
- One2One:
- One2Seq:



- Deep keyphrase generation. R Meng, S Zhao, S Han, D He, P Brusilovsky, Y Chi, 2017.
- One size does not fit all: Generating and evaluating variable number of keyphrases. X Yuan, T Wang, R Meng, K Thaker, P Brusilovsky, D He, A Trischler, 2020.

Seq2Seq for KPG: One2One vs One2Seq

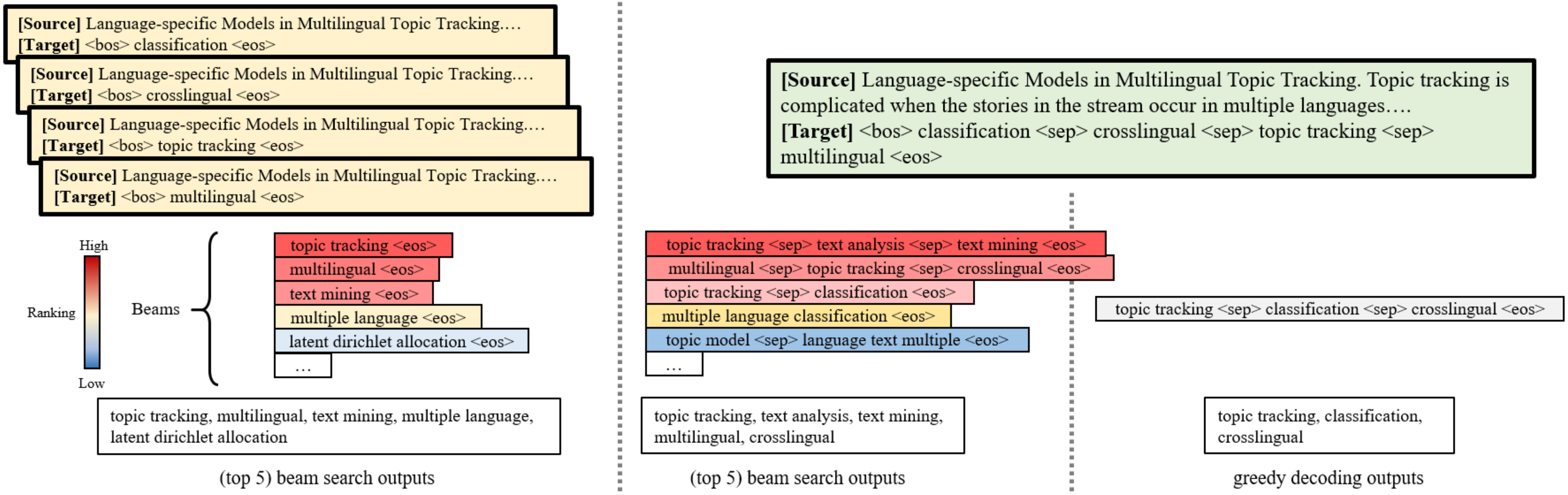
- One2One:
- One2Seq:



- Deep keyphrase generation. R Meng, S Zhao, S Han, D He, P Brusilovsky, Y Chi, 2017.
- One size does not fit all: Generating and evaluating variable number of keyphrases. X Yuan, T Wang, R Meng, K Thaker, P Brusilovsky, D He, A Trischler, 2020.

Seq2Seq for KPG: One2One vs One2Seq

- One2One:
- One2Seq:



- Deep keyphrase generation. R Meng, S Zhao, S Han, D He, P Brusilovsky, Y Chi, 2017.
- One size does not fit all: Generating and evaluating variable number of keyphrases. X Yuan, T Wang, R Meng, K Thaker, P Brusilovsky, D He, A Trischler, 2020.

Research Questions

- Q1: How well do KPG models generalize to various testing distributions?
- Q2: Does the order of target keyphrases matter while training One2Seq?
- Q3: Does more training data help? How to better make use of them?
- Q4: Is copy mechanism always helpful for KPG models?
- Q5: What is the effect of beam width?

Research Questions

- **Q1: How well do KPG models generalize to various testing distributions?**
- **Q2: Does the order of target keyphrases matter while training One2Seq?**
- **Q3: Does more training data help? How to better make use of them?**
- Q4: Is copy mechanism always helpful for KPG models?
  - No
- Q5: What is the effect of beam width?
  - For now, the larger the better

Q1: How well do various KPG models generalize?

Training Paradigm

- One2One vs. One2Seq

Architecture

RNN vs. Transformer

Q1: How well do various KPG models generalize?

	Dataset
D_0	KP20K
	KRAPIVIN
	D_0 Average
D_1	INSPEC
	NUS
	SEMEVAL
	D_1 Average
D_2	DUC
All	Average

Dataset

D_0 In-distribution

D_1 Out-of-distribution

D_2 Out-of-domain

Training Paradigm

- One2One vs. One2Seq

Architecture

- RNN vs. Transformer

Q1: How well do various KPG models generalize?

	Dataset	Present ($F_1@\mathcal{O}$)				Absent (**R@50**)			
		One2One		One2Seq		One2One		One2Seq	
		RNN	TRANS	RNN	TRANS	RNN	TRANS	RNN	TRANS
D_0	KP20K	35.3	37.4	31.2	36.2	13.1	22.1	3.2	15.0
	KRAPIVIN	35.5	33.0	33.5	36.4	13.7	23.8	3.3	16.6
	D_0 Average	35.4	35.2	32.3	**36.3**	13.4	**23.0**	3.2	15.8
D_1	INSPEC	33.7	32.6	38.8	36.9	8.2	9.2	3.7	6.7
	NUS	43.4	41.1	39.2	42.3	11.2	18.9	2.9	12.5
	SEMEVAL	35.2	35.1	36.2	34.8	6.1	18.9	1.7	12.5
	D_1 Average	37.4	36.3	**38.1**	38.0	8.5	**12.7**	2.8	9.2
D_2	DUC	13.4	7.8	**15.0**	11.0	0.0	**0.2**	0.0	0.0
All	Average	32.8	31.2	32.3	**32.9**	8.7	**14.0**	2.5	9.8

Q1: How well do various KPG models generalize?

	Dataset	Present ($F_1@\mathcal{O}$)				Absent (**R@50**)			
		One2One		One2Seq		One2One		One2Seq	
		RNN	TRANS	RNN	TRANS	RNN	TRANS	RNN	TRANS
D_0	KP20K	35.3	37.4	31.2	36.2	13.1	22.1	3.2	15.0
	KRAPIVIN	35.5	33.0	33.5	36.4	13.7	23.8	3.3	16.6
	D_0 Average	35.4	35.2	32.3	**36.3**	13.4	**23.0**	3.2	15.8
D_1	INSPEC	33.7	32.6	38.8	36.9	8.2	9.2	3.7	6.7
	NUS	43.4	41.1	39.2	42.3	11.2	18.9	2.9	12.5
	SEMEVAL	35.2	35.1	36.2	34.8	6.1	18.9	1.7	12.5
	D_1 Average	37.4	36.3	**38.1**	38.0	8.5	**12.7**	2.8	9.2
D_2	DUC	13.4	7.8	**15.0**	11.0	0.0	**0.2**	0.0	0.0
All	Average	32.8	31.2	32.3	**32.9**	8.7	**14.0**	2.5	9.8

<

- **One2Seq** generalizes better on present KPG

Q1: How well do various KPG models generalize?

	Dataset	Present ($F_1@\mathcal{O}$)				Absent (**R@50**)			
		One2One		One2Seq		One2One		One2Seq	
		RNN	TRANS	RNN	TRANS	RNN	TRANS	RNN	TRANS
D_0	KP20K	35.3	37.4	31.2	36.2	13.1	22.1	3.2	15.0
	KRAPIVIN	35.5	33.0	33.5	36.4	13.7	23.8	3.3	16.6
	D_0 Average	35.4	35.2	32.3	**36.3**	13.4	**23.0**	3.2	15.8
D_1	INSPEC	33.7	32.6	38.8	36.9	8.2	9.2	3.7	6.7
	NUS	43.4	41.1	39.2	42.3	11.2	18.9	2.9	12.5
	SEMEVAL	35.2	35.1	36.2	34.8	6.1	18.9	1.7	12.5
	D_1 Average	37.4	36.3	**38.1**	38.0	8.5	**12.7**	2.8	9.2
D_2	DUC	13.4	7.8	**15.0**	11.0	0.0	**0.2**	0.0	0.0
All	Average	32.8	31.2	32.3	**32.9**	8.7	**14.0**	2.5	9.8

>

- One2Seq generalizes better on present KPG, while One2One excels at absent KPG.

Q1: How well do various KPG models generalize?

	Dataset	Present ($F_1@\mathcal{O}$)				Absent (**R@50**)			
		One2One		One2Seq		One2One		One2Seq	
		RNN	TRANS	RNN	TRANS	RNN	TRANS	RNN	TRANS
D_0	KP20K	$\boxed{35.3}$ <	$\boxed{37.4}$	$\boxed{31.2}$ <	$\boxed{36.2}$	13.1	22.1	3.2	15.0
	KRAPIVIN	35.5	33.0	33.5	36.4	13.7	23.8	3.3	16.6
	D_0 Average	35.4	35.2	32.3	**36.3**	13.4	**23.0**	3.2	15.8
D_1	INSPEC	33.7	32.6	38.8	36.9	8.2	9.2	3.7	6.7
	NUS	43.4	41.1	39.2	42.3	11.2	18.9	2.9	12.5
	SEMEVAL	35.2	35.1	36.2	34.8	6.1	18.9	1.7	12.5
	D_1 Average	37.4	36.3	**38.1**	38.0	8.5	**12.7**	2.8	9.2
D_2	DUC	13.4	7.8	**15.0**	11.0	0.0	**0.2**	0.0	0.0
All	Average	32.8	31.2	32.3	**32.9**	$\boxed{8.7}$ <	$\boxed{14.0}$	$\boxed{2.5}$ <	$\boxed{9.8}$

- Transformer fits better on in-distribution data and exhibits much better abstractiveness

Q1: How well do various KPG models generalize?

	Dataset	Present ($F_1@\mathcal{O}$)				Absent (R@50)			
		One2One		One2Seq		One2One		One2Seq	
		RNN	TRANS	RNN	TRANS	RNN	TRANS	RNN	TRANS
D_0	KP20K	35.3	37.4	31.2	36.2	13.1	22.1	3.2	15.0
	KRAPIVIN	35.5	33.0	33.5	36.4	13.7	23.8	3.3	16.6
	D_0 Average	35.4	35.2	32.3	**36.3**	13.4	**23.0**	3.2	15.8
D_1	INSPEC	33.7	32.6	38.8	36.9	8.2	9.2	3.7	6.7
	NUS	43.4	41.1	39.2	42.3	11.2	18.9	2.9	12.5
	SEMEVAL	35.2	35.1	36.2	34.8	6.1	18.9	1.7	12.5
	D_1 Average	37.4	36.3	**38.1**	38.0	8.5	**12.7**	2.8	9.2
D_2	DUC	13.4	7.8	**15.0**	11.0	0.0	**0.2**	0.0	0.0
All	Average	32.8	31.2	32.3	**32.9**	8.7	**14.0**	2.5	9.8

- Transformer fits better on in-distribution data and exhibits much better abstractiveness
- But RNN seems to generalize better on out-of-distribution present KPG

Q1: How well do various KPG models generalize?

- **Training Paradigm:** **One2One or One2Seq ?**
- **Architecture:** **RNN vs. Transformer ?**

Q1: How well do various KPG models generalize?

- **Training Paradigm:** One2One or One2Seq ?
- **Architecture:** RNN vs. Transformer ?

- **It depends!**
  - Prefer present? Transformer + One2Seq
  - Prefer absent? Transformer + One2One
  - Less computational resources? RNN + One2One

Q2: Does order matter?

[Source] Language-specific Models in Multilingual Topic Tracking. Topic tracking is complicated when the stories in the stream occur in multiple languages….
[Target] <bos> classification <sep> crosslingual <sep> topic tracking <sep> multilingual <eos>

For training One2Seq models, we can concatenate target keyphrases in different orders:

- Alpha (A->Z) / Alpha-rev (Z->A)
- Short -> Long / Long -> Short
- Original / Original-rev
- Present-Absent / Absent-Present
- Random

Q2: Does order matter?

[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. …

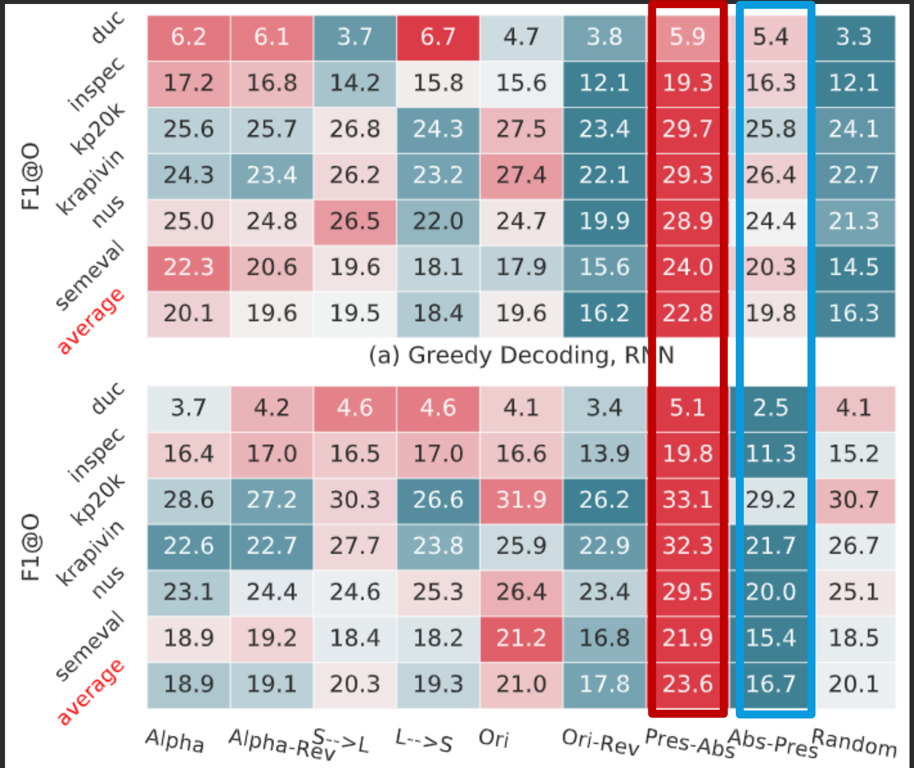
[Target Sequence]=keyphrases
[classification, crosslingual, Arabic, TDT, topic tracking, multilingual]

[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
Original	**[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
Abs-Pres	**[Source]** Language-specific Models in Multilingual Topic **[Target]** <bos> Arabic <sep> TDT <sep>classification <sep> crosslingual <sep> multilingual <sep> topic tracking
Pres-Abs	**[Source]** Language-specific Models in Multilingual Topic **[Target]** <bos> multilingual <sep> topic tracking <sep> TDT <sep> Arabic <sep> classification <sep> crosslingual

Q2: Does order matter?

[Source] Language-specific Models in Multilingual Topic Tracking. Topic tracking is complicated when the stories in the stream occur in multiple languages….
[Target] <bos> classification <sep> crosslingual <sep> topic tracking <sep> multilingual <eos>



(a) Greedy Decoding, RNN

(c) Greedy Decoding, Transformer

- With greedy decoding, target phrase order shows distinct effects on performance (i.e. Pres-Abs >> Abs-Pres).

Q2: Does order matter?

[Source] Language-specific Models in Multilingual Topic Tracking. Topic tracking is complicated when the stories in the stream occur in multiple languages.…
[Target] <bos> classification <sep> crosslingual <sep> topic tracking <sep> multilingual <eos>

F1@O	Alpha	Alpha-Rev	S-->L	L-->S	Ori	Ori-Rev	Pres-Abs	Abs-Pres	Random
duc	6.2	6.1	3.7	6.7	4.7	3.8	5.9	5.4	3.3
inspec	17.2	16.8	14.2	15.8	15.6	12.1	19.3	16.3	12.1
kp20k	25.6	25.7	26.8	24.3	27.5	23.4	29.7	25.8	24.1
krapivin	24.3	23.4	26.2	23.2	27.4	22.1	29.3	26.4	22.7
nus	25.0	24.8	26.5	22.0	24.7	19.9	28.9	24.4	21.3
semeval	22.3	20.6	19.6	18.1	17.9	15.6	24.0	20.3	14.5
average	20.1	19.6	19.5	18.4	19.6	16.2	22.8	19.8	16.3

(a) Greedy Decoding, RNN

F1@O	Alpha	Alpha-Rev	S-->L	L-->S	Ori	Ori-Rev	Pres-Abs	Abs-Pres	Random
duc	15.4	15.6	16.6	17.2	15.9	16.7	15.0	12.9	14.5
inspec	37.4	36.5	37.4	38.5	40.7	36.9	38.8	37.6	37.2
kp20k	33.5	32.9	33.6	32.8	34.4	33.6	31.2	32.7	33.9
krapivin	34.3	34.7	34.6	34.5	36.4	34.5	33.5	34.3	34.2
nus	42.5	41.1	40.0	41.7	41.8	40.6	39.2	41.0	42.6
semeval	35.5	36.2	37.2	36.5	36.3	35.7	36.2	34.6	35.9
average	33.1	32.8	33.3	33.5	34.3	33.0	32.3	32.2	33.0

(b) Beam Size 50, RNN

F1@O	Alpha	Alpha-Rev	S-->L	L-->S	Ori	Ori-Rev	Pres-Abs	Abs-Pres	Random
duc	3.7	4.2	4.6	4.6	4.1	3.4	5.1	2.5	4.1
inspec	16.4	17.0	16.5	17.0	16.6	13.9	19.8	11.3	15.2
kp20k	28.6	27.2	30.3	26.6	31.9	26.2	33.1	29.2	30.7
krapivin	22.6	22.7	27.7	23.8	25.9	22.9	32.3	21.7	26.7
nus	23.1	24.4	24.6	25.3	26.4	23.4	29.5	20.0	25.1
semeval	18.9	19.2	18.4	18.2	21.2	16.8	21.9	15.4	18.5
average	18.9	19.1	20.3	19.3	21.0	17.8	23.6	16.7	20.1

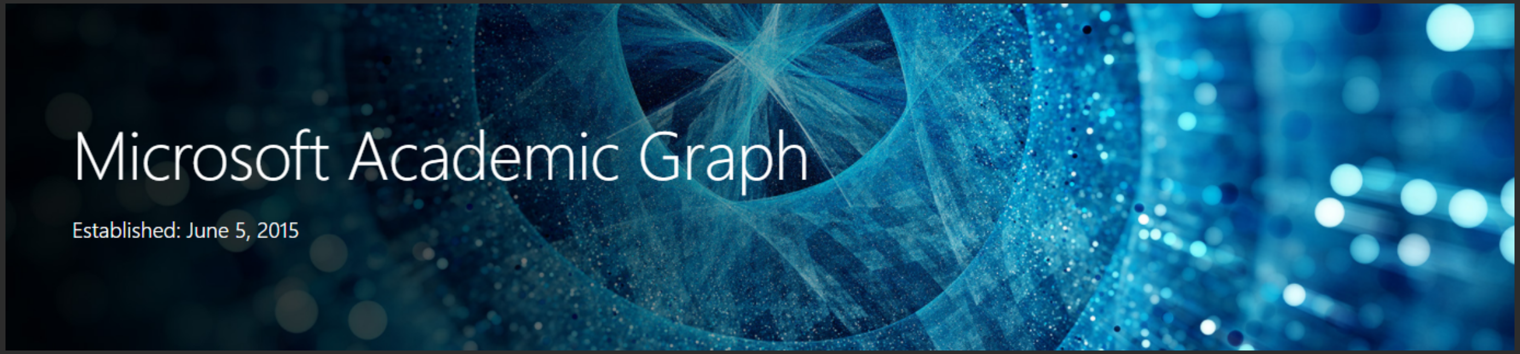
(c) Greedy Decoding, Transformer

F1@O	Alpha	Alpha-Rev	S-->L	L-->S	Ori	Ori-Rev	Pres-Abs	Abs-Pres	Random
duc	10.9	13.0	11.4	12.2	9.5	13.4	11.0	10.5	10.0
inspec	35.0	35.6	35.9	36.2	36.1	37.4	36.9	34.7	35.4
kp20k	35.2	35.5	34.9	35.5	36.1	35.5	36.2	35.4	35.8
krapivin	34.1	36.4	34.6	34.7	34.1	35.5	36.4	35.9	34.7
nus	41.9	44.5	41.5	42.4	41.5	42.5	42.3	40.8	41.1
semeval	33.9	34.1	33.7	35.3	36.7	34.4	34.8	34.5	35.7
average	31.8	33.2	32.0	32.7	32.3	33.1	32.9	32.0	32.1

(d) Beam Size 50, Transformer

- The effect of target ordering diminishes when beam search is performed, especially with large beam size.

Q3: Extra (noisy) data?

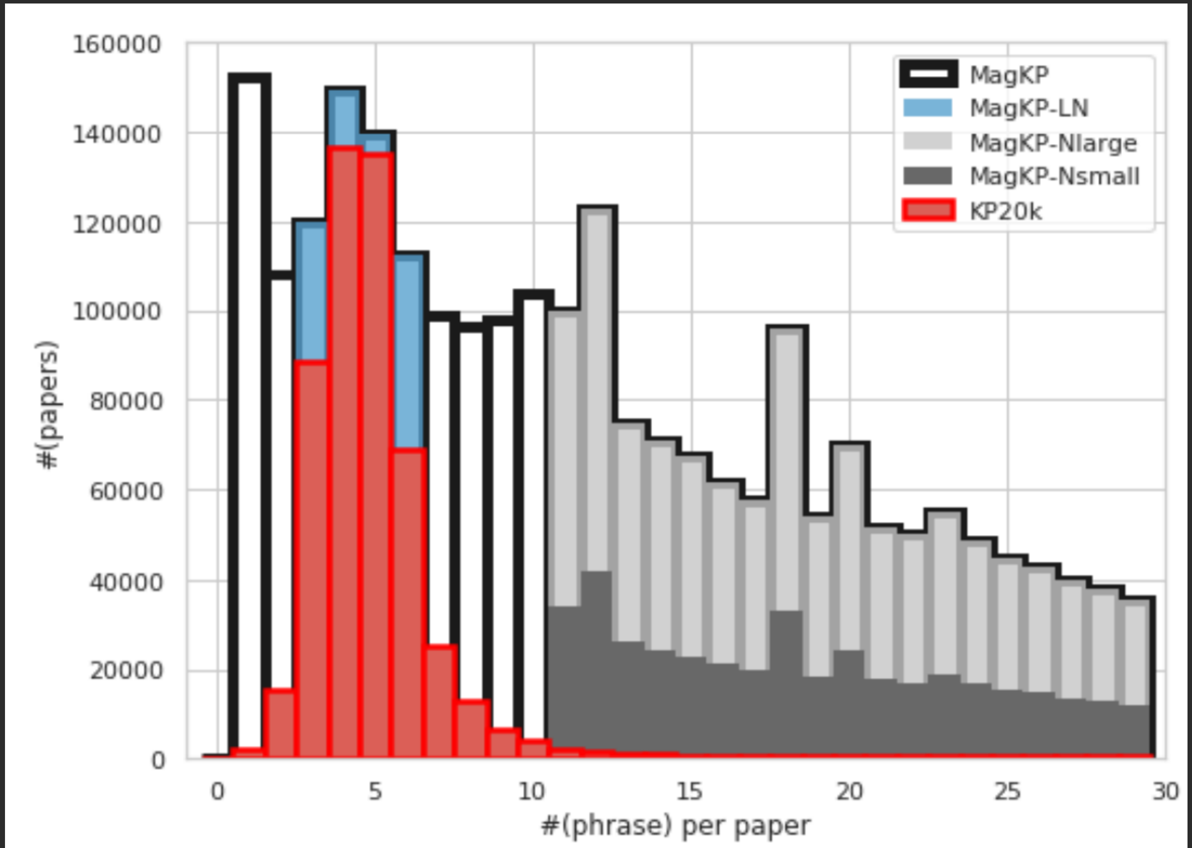


MagKP dataset (2.7M)

5x larger than KP20k (514K)

Q3: Extra (noisy) data?

- MagKP is also noisy
- Distribution of MagKP is very different from normal author-keyword datasets e.g. KP20k
  - Authors usually provide 3~10 keyphrases for a paper, **μ**=5.25 (red-filled bars)
  - MagKP can have up to 100 keyphrases for a paper, **μ**=15.4 (black-bordered bars)

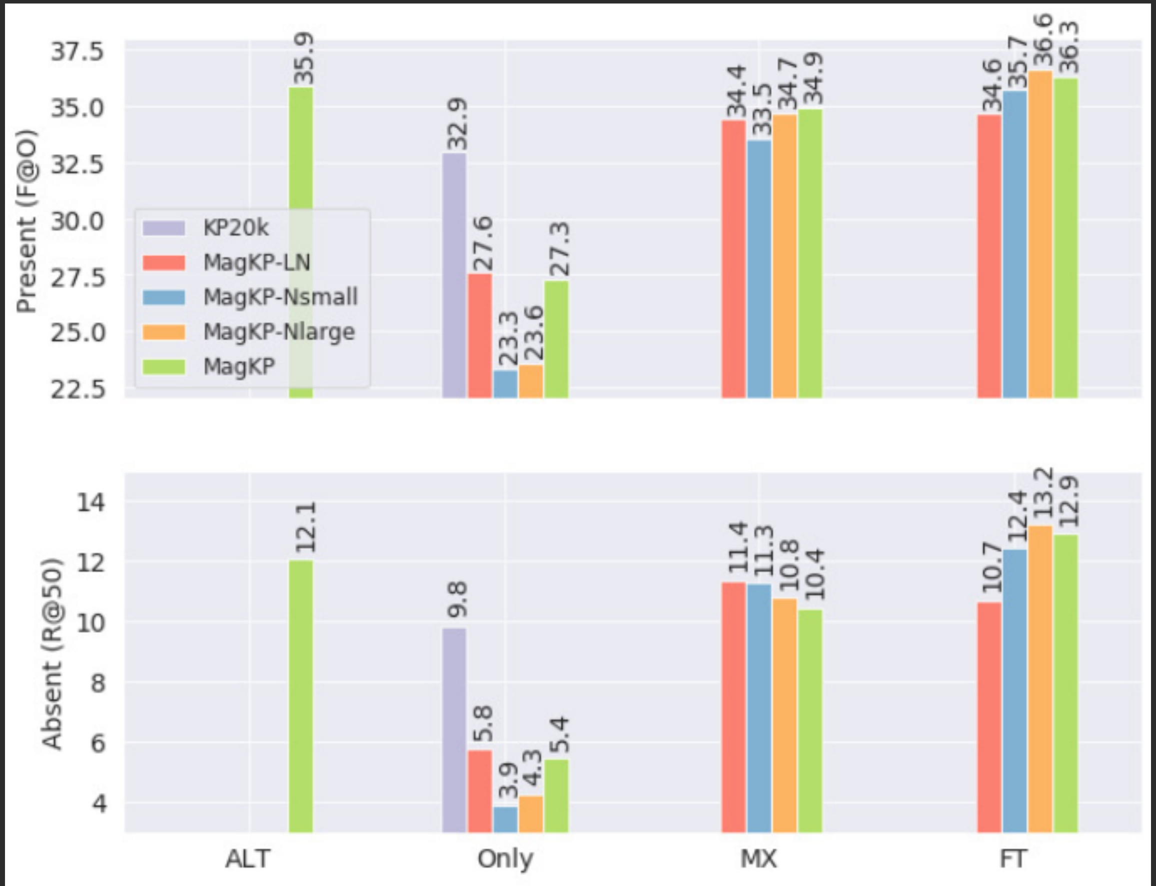


Q3: Extra (noisy) data?

- Extra data does help
- Transformer + One2Seq achieves SOTA present scores

Q3: Extra (noisy) data?

- Pre-training w/ noisier data performs better.
- The way of mixing noisy/clean data also makes a difference
  - FT (Fine-Tuning) > ALT > MX >> Only
  - Pre-training w/ noisy data and then fine-tuning w/ clean data can lead to better performance



	Kp20K			Krapivin			Inspec			NUS			SemEval		
Model	5	10	$\mathcal{O}$	5	10	$\mathcal{O}$	5	10	$\mathcal{O}$	5	10	$\mathcal{O}$	5	10	$\mathcal{O}$
`One2One variants`															
RNN-O2O-KP20k	33.1	27.9	35.3	32.0	27.0	35.5	28.5	32.5	33.7	40.2	35.9	43.4	32.9	34.6	35.2
RNN-O2O-KP20k-nocopy	8.3	8.5	8.4	5.5	5.8	5.0	4.9	5.3	5.3	10.3	10.9	10.4	8.4	8.4	7.5
RNN-O2O-KP20k+MagKP-ALT	32.4	27.4	34.7	32.3	28.1	35.6	32.2	37.6	38.4	40.2	38.2	43.5	35.4	35.4	37.4
BIGRNN-O2O-KP20k	35.5	29.5	38.1	34.2	29.1	38.8	31.0	36.2	37.3	42.6	39.2	45.8	34.4	36.1	37.1
BIGRNN-O2O-magkp20k-ALT	33.1	27.9	35.3	32.3	28.1	36.4	32.2	37.1	37.8	41.2	38.1	44.9	35.6	35.7	36.0
TF-O2O-KP20k	34.5	28.9	37.4	29.5	26.4	33.0	28.0	30.8	32.6	37.6	36.1	41.1	32.9	33.0	35.1
TF-O2O-KP20k-nocopy	28.5	24.0	31.2	24.7	20.8	28.9	16.4	18.1	18.4	33.8	30.2	35.5	25.3	25.9	25.8
TF-O2O-KP20k+MagKP-ALT	33.8	28.2	36.5	32.7	27.6	34.9	28.9	33.5	34.0	39.3	37.8	41.9	33.7	34.8	35.7
`One2Seq variants`															
RNN-O2S-KP20k	31.2	26.1	31.2	30.9	26.9	33.5	32.8	38.7	38.8	37.3	36.6	39.2	33.5	35.0	36.2
RNN-O2S-KP20k-nocopy	10.4	10.2	11.1	8.1	7.9	9.8	4.4	4.5	4.5	11.0	10.6	11.0	8.8	8.6	8.8
RNN-O2S+KP20k+MagKP-ALT	28.2	23.8	28.2	28.0	25.8	30.6	32.9	40.3	39.9	35.1	33.2	36.4	30.6	33.1	34.0
BIGRNN-O2S-KP20k	30.2	25.7	30.4	29.8	26.4	32.4	31.6	37.5	38.1	37.4	35.7	39.7	32.5	33.7	35.3
BIGRNN-O2S-KP20k+MagKP-ALT	28.2	23.7	28.2	28.9	25.6	30.9	34.9	41.1	40.1	35.9	34.3	37.6	32.0	33.8	34.8
TF-O2S-KP20k	34.6	29.0	36.2	32.4	28.1	36.4	31.5	36.6	36.9	40.1	37.3	42.3	33.9	34.2	34.8
TF-O2S-KP20k-nocopy	32.3	29.0	33.9	28.5	25.1	31.5	23.2	24.6	25.3	36.9	34.5	37.5	27.4	28.4	29.5
TF-O2S-KP20k+MagKP-ALT	36.8	30.2	37.7	35.2	29.9	37.6	32.2	38.8	39.4	41.8	39.2	44.1	35.6	36.5	38.7
`TRANS+One2Seq`															
MagKP-LN-ONLY	28.1	25.1	28.0	27.8	26.4	28.7	29.6	34.3	34.3	33.5	34.0	34.9	28.9	30.3	30.2
MagKP-Nsmall-ONLY	20.8	19.8	20.9	25.2	24.3	26.0	30.8	34.0	33.9	26.2	27.0	27.0	24.1	26.2	24.8
MagKP-Nlarge-ONLY	20.4	19.6	21.1	24.8	23.5	25.6	32.6	36.2	36.1	26.0	26.6	28.1	21.4	25.0	23.3
MagKP-ONLY	25.3	22.3	25.5	26.2	25.1	28.0	31.3	38.7	37.2	29.9	31.0	31.3	26.5	30.3	29.5
MagKP-LN-MX	35.5	29.3	36.9	34.2	28.6	37.9	31.5	38.0	37.8	41.7	38.7	44.6	32.7	35.0	34.5
MagKP-Nsmall-MX	35.3	29.2	36.5	34.1	28.7	37.0	31.6	38.2	37.2	40.6	38.5	42.6	33.4	36.2	35.5
MagKP-Nlarge-MX	36.3	30.0	37.1	34.9	29.7	36.9	31.8	37.8	38.3	41.9	39.5	44.8	34.4	35.2	37.6
MagKP-MX	36.3	29.8	37.4	35.0	30.0	37.3	32.0	39.8	38.7	41.4	39.8	44.6	34.8	36.7	36.9
MagKP-LN-FT	36.2	29.8	37.7	34.9	29.6	36.2	32.1	38.0	38.2	41.4	39.4	44.9	35.4	36.2	36.8
MagKP-Nsmall-FT	36.4	30.1	37.7	35.7	30.5	39.4	32.6	38.4	38.8	43.0	39.5	45.6	34.9	35.6	37.4
MagKP-Nlarge-FT	37.0	30.4	37.9	36.6	30.6	38.9	33.3	39.5	39.8	44.0	40.2	47.9	34.3	36.4	35.9
MagKP-FT	37.1	30.5	38.3	36.1	30.6	38.4	32.4	38.1	38.5	43.9	40.1	46.0	36.6	37.0	39.2
`Abstractive Neural Generation`															
SotaMax	36.0	29.8	35.7	32.9	28.5	37.1	29.6	35.7	33.1	37.6	36.6	40.6	32.7	35.2	35.7

Conclusion

- Basic settings are critical
  - Our study provides a guideline on how to choose such settings

- Open questions
  - More efficient KPG inference
  - Mitigate the effect of phrase ordering
  - Better way utilizing large and noisy data

Thank you!

arXiv: 2009.10229
Code & Data: https://github.com/memray/OpenNMT-kpg-release

rui.meng@pitt.edu
eric.yuan@microsoft.com
tong.wang@microsoft.com