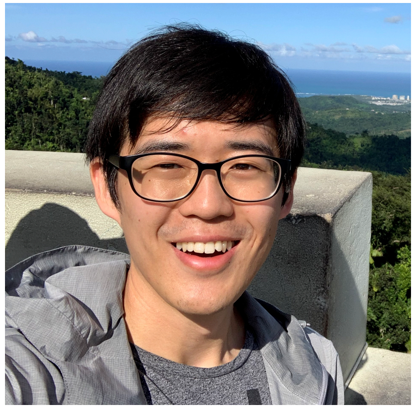
ONE SIZE DOES NOT FIT ALL: GENERATING AND EVALUATING VARIABLE NUMBER OF KEYPHRASES

Xingdi Yuan*

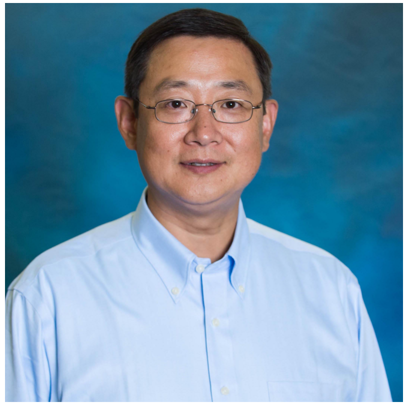
Tong Wang*



Rui Meng*

Khushboo Thaker

Peter Brusilovsky

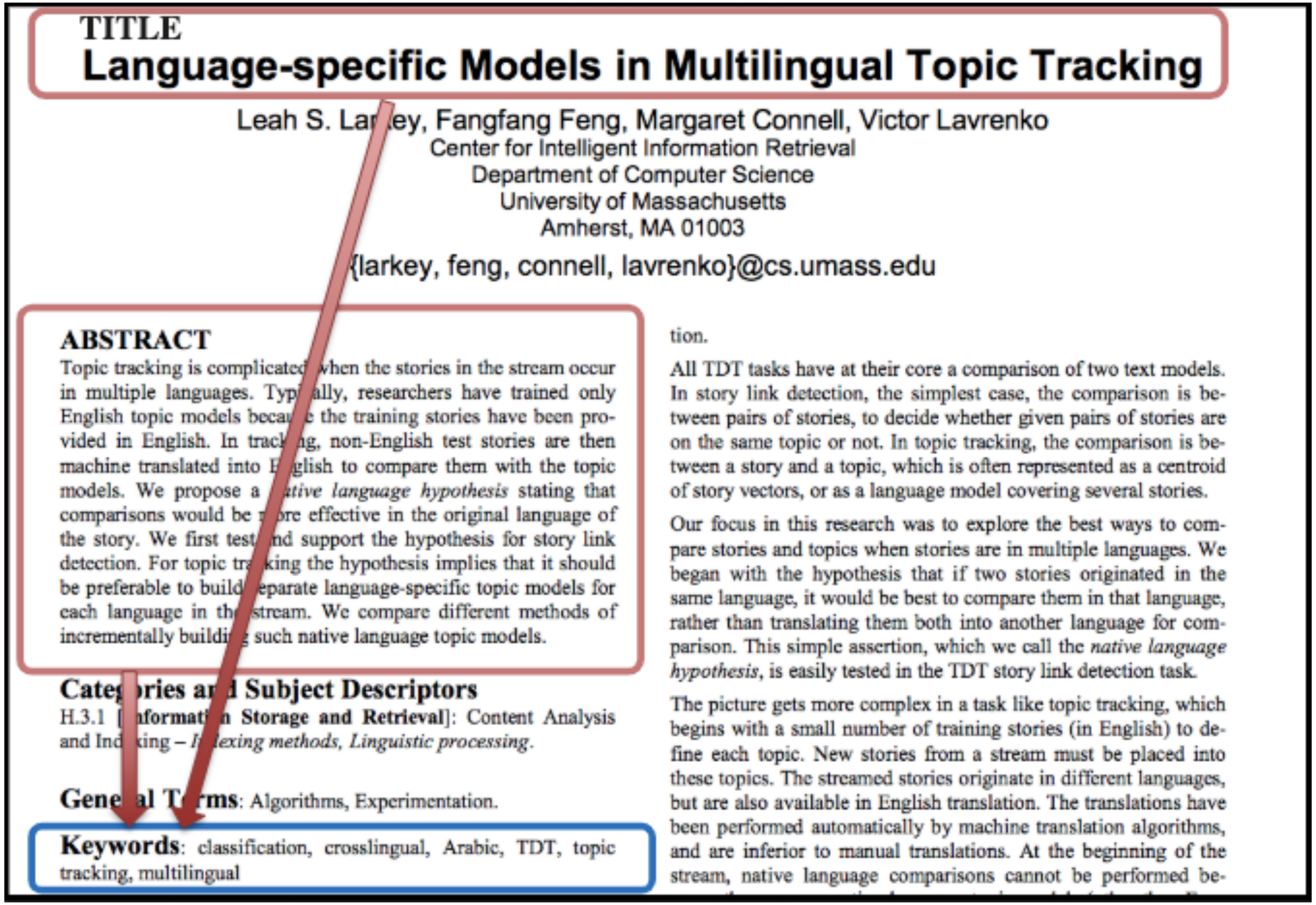


Daqing He

Adam Trischler

* Equal contribution. The order is determined by a fidget spinner.

What is keyphrase generation?



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.

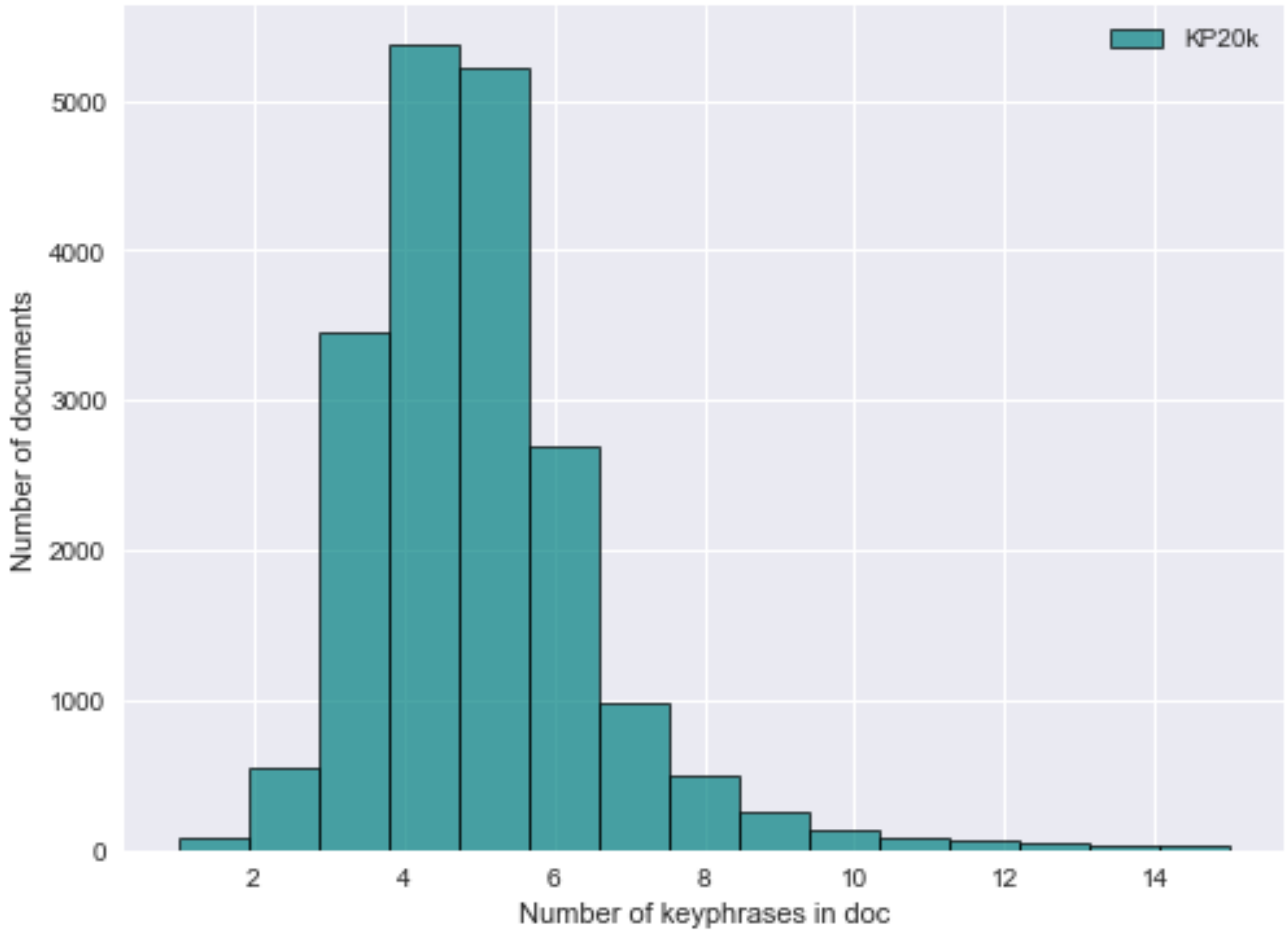
Keywords: classification, crosslingual, Arabic, TDT, topic tracking, multilingual

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.

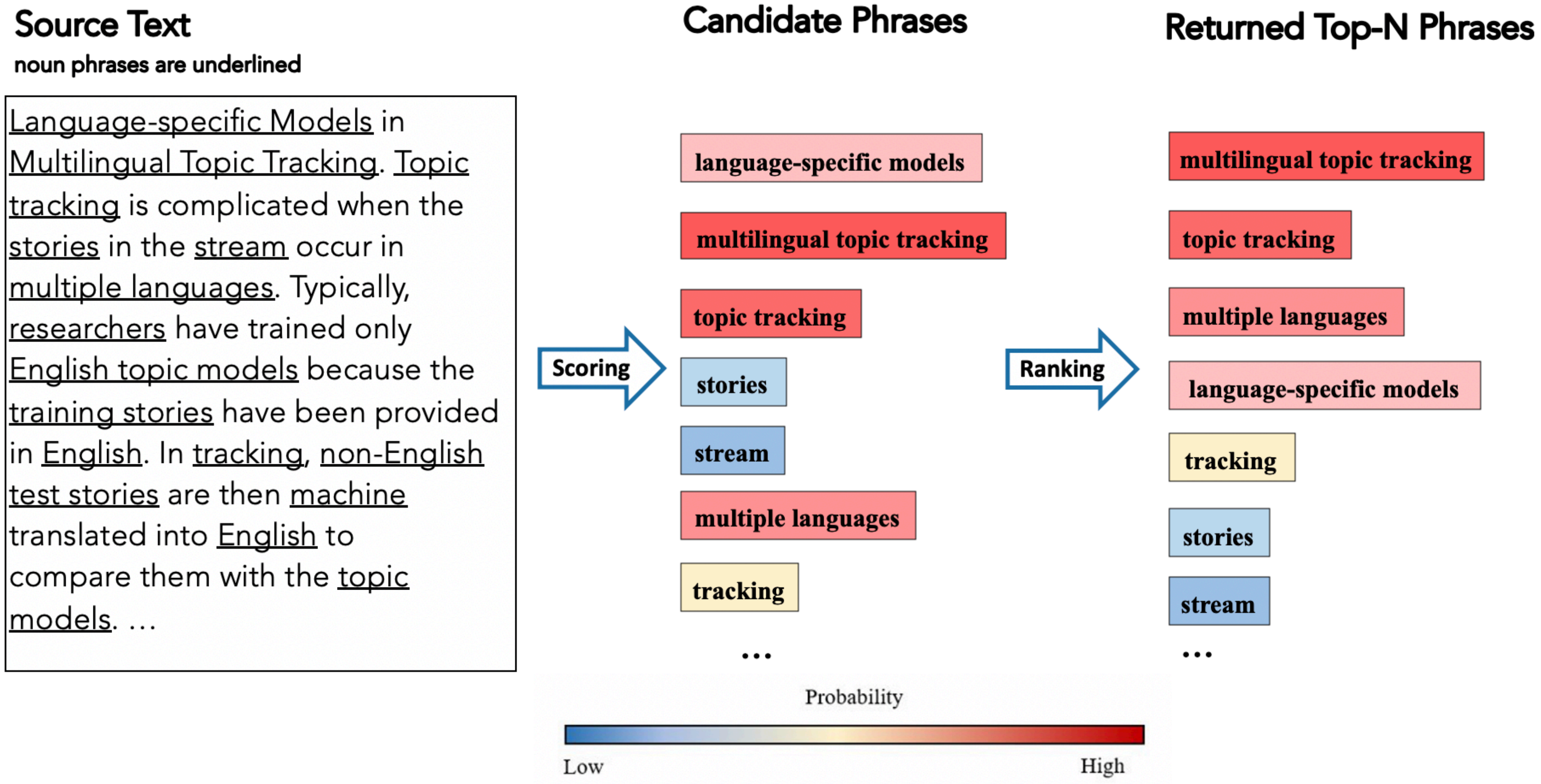
The picture gets more complex in a task like topic tracking, which begins with a small number of training stories (in English) to define each topic. New stories from a stream must be placed into these topics. The streamed stories originate in different languages, but are also available in English translation. 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-



By nature, the number of keyphrases of a document is variable.

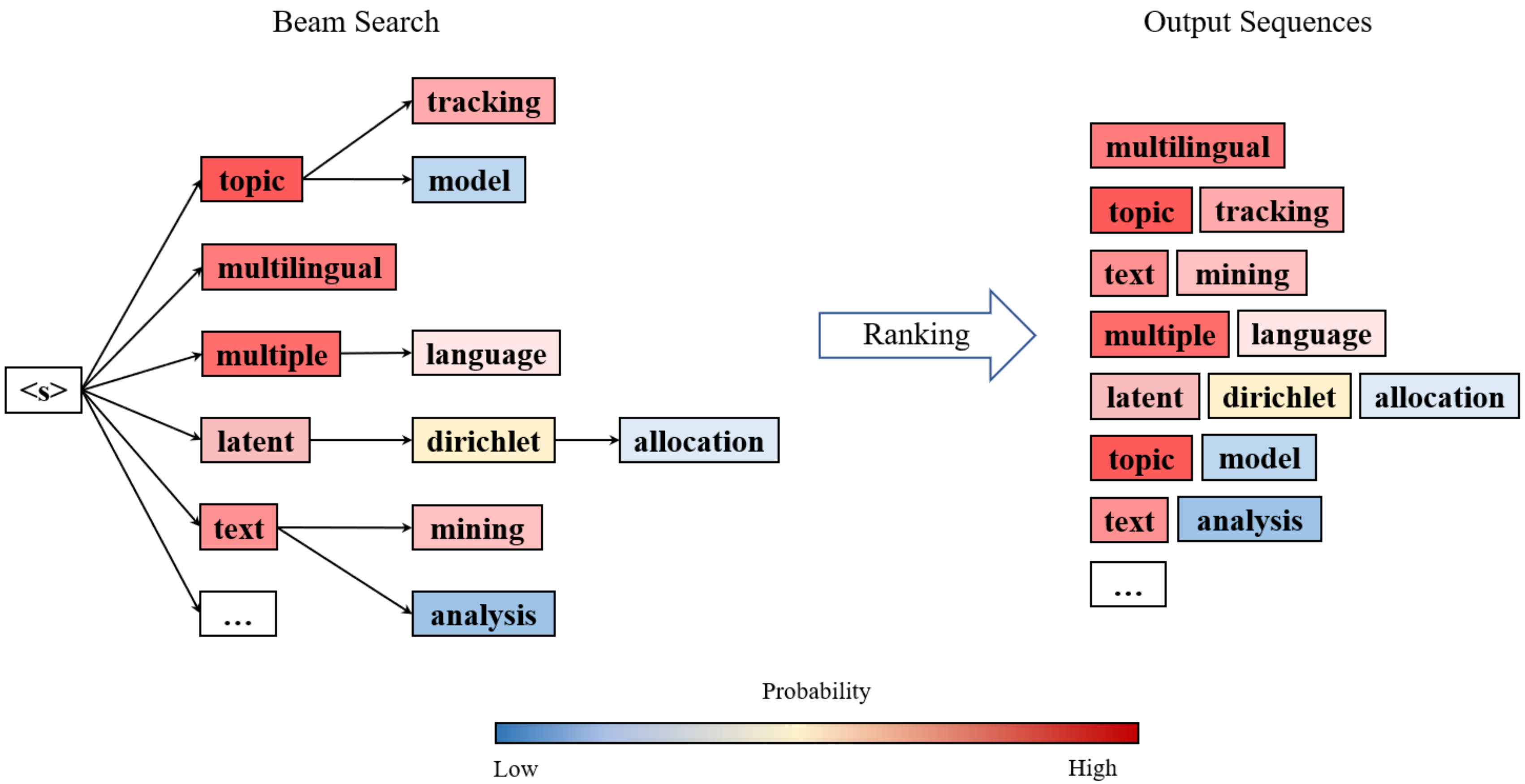
Previous works

- Extractive models rank a long candidate list, e.g. noun phrases, n-grams



Previous works

- Prior generative studies use beam search to output many phrases.



Previous evaluation metrics

$$P@k = \frac{\boxed{|\hat{\mathcal{Y}}_{:k} \cap \mathcal{Y}|}}{\boxed{|\hat{\mathcal{Y}}_{:k}|}}$$

Correct predictions

Model predictions

$$R@k = \frac{|\hat{\mathcal{Y}}_{:k} \cap \mathcal{Y}|}{\boxed{|\mathcal{Y}|}}$$

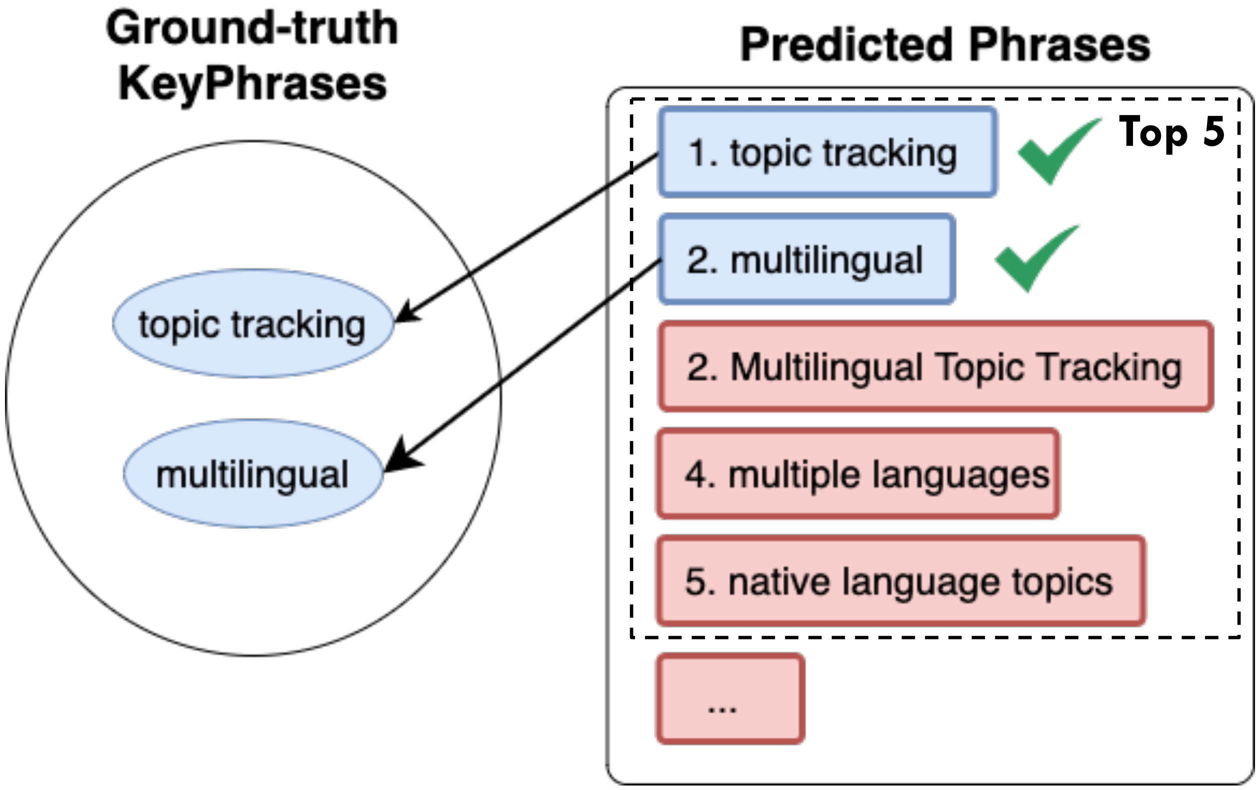
Gold standard

$$F_1@k = \frac{2 \times P@k \times R@k}{P@k + R@k}$$

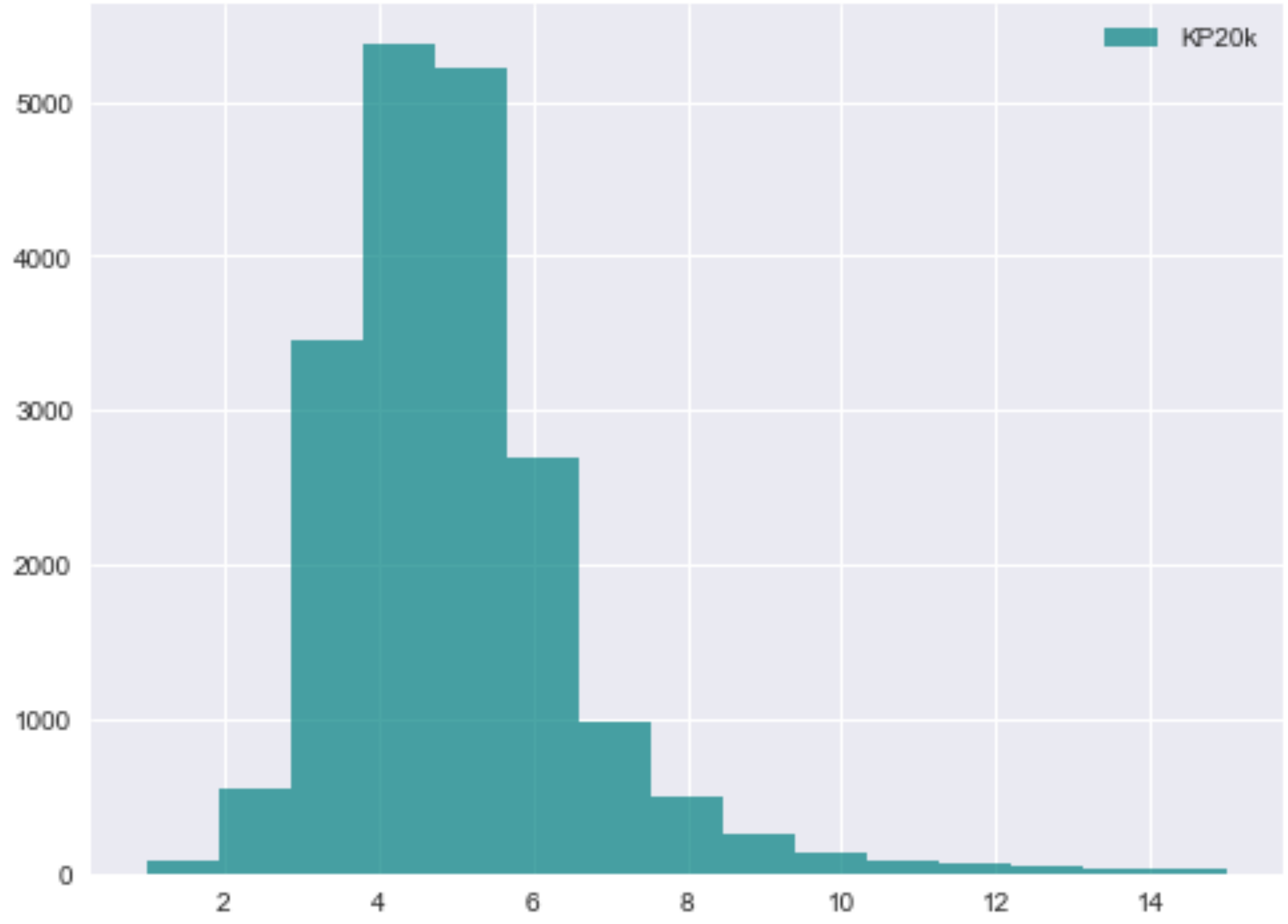
Where $\hat{\mathcal{Y}}_{:k}$ are a model's top k predictions.

k is typically 5 or 10.

Imagine an oracle model...



P@5=0.4, R@5=1.0, F1@5=0.571



On KP20K dataset, the performance upper bound for a ranking-based oracle model is $F_1@5=0.858$ and $F_1@10=0.626$.

Generating and Evaluating Variable Number of Keyphrases

CopyRNN (Meng et al., 2017)

[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 </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>

CatSeq: generating the concatenation of keyphrases

[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. …

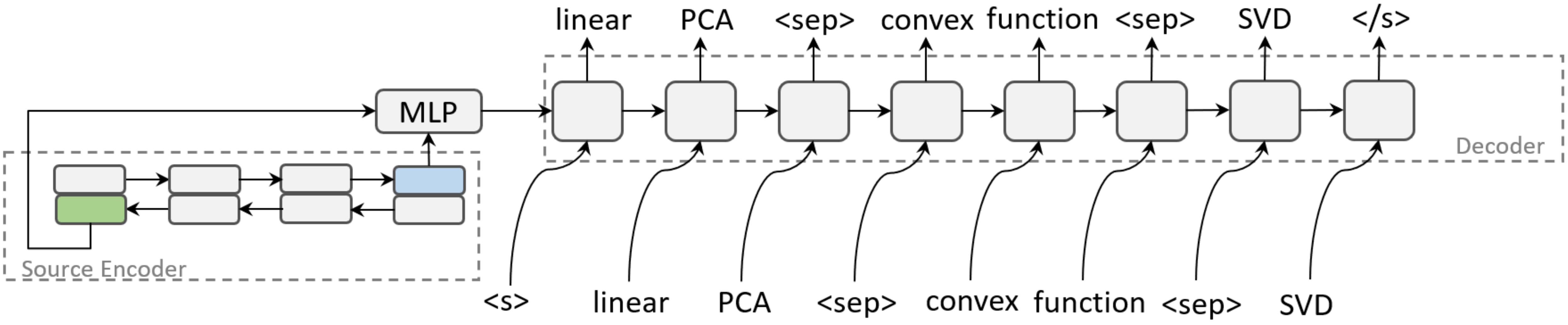
[Target Sequence]

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

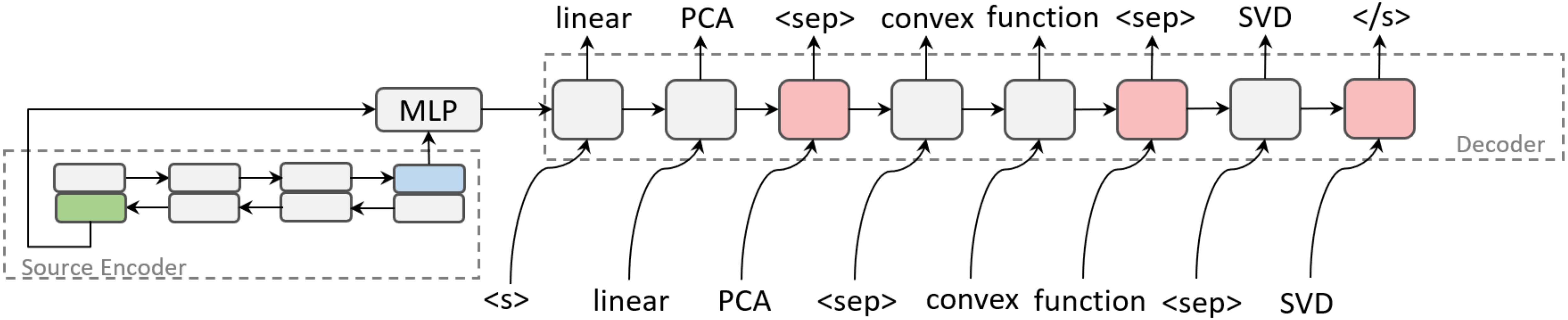
[Source] Language-specific Models in Multilingual Topic Tracking.…

[Target] <s> classification <sep> crosslingual <sep> Arabic <sep> TDT <sep> topic tracking <sep> multilingual </s>

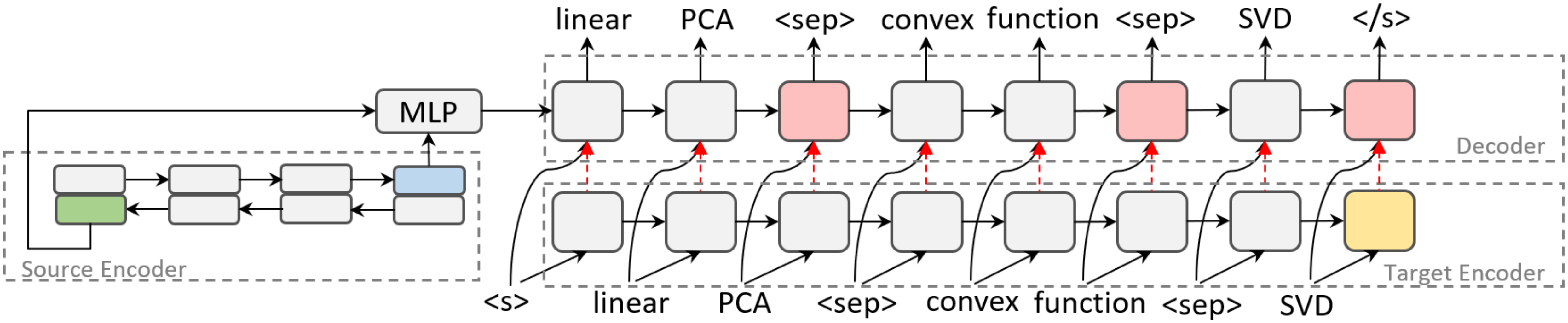
CatSeq: generating the concatenation of keyphrases

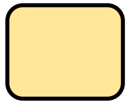


CatSeq + Orthogonal Regularization



CatSeqD = CatSeq + Orthogonal Regularization + Semantic Coverage



Maximizing the mutual information between  and 

Evaluating Variable Number of Keyphrases

$$P@k = \frac{|\hat{\mathcal{Y}}_{:k} \cap \mathcal{Y}|}{|\hat{\mathcal{Y}}_{:k}|}$$

Correct predictions

Model predictions

$$R@k = \frac{|\hat{\mathcal{Y}}_{:k} \cap \mathcal{Y}|}{|\mathcal{Y}|}$$

Gold standard

$$F_1@k = \frac{2 \times P@k \times R@k}{P@k + R@k}$$

$\mathbf{F}_1@\mathcal{O} \; : \; k = |\mathcal{Y}|$

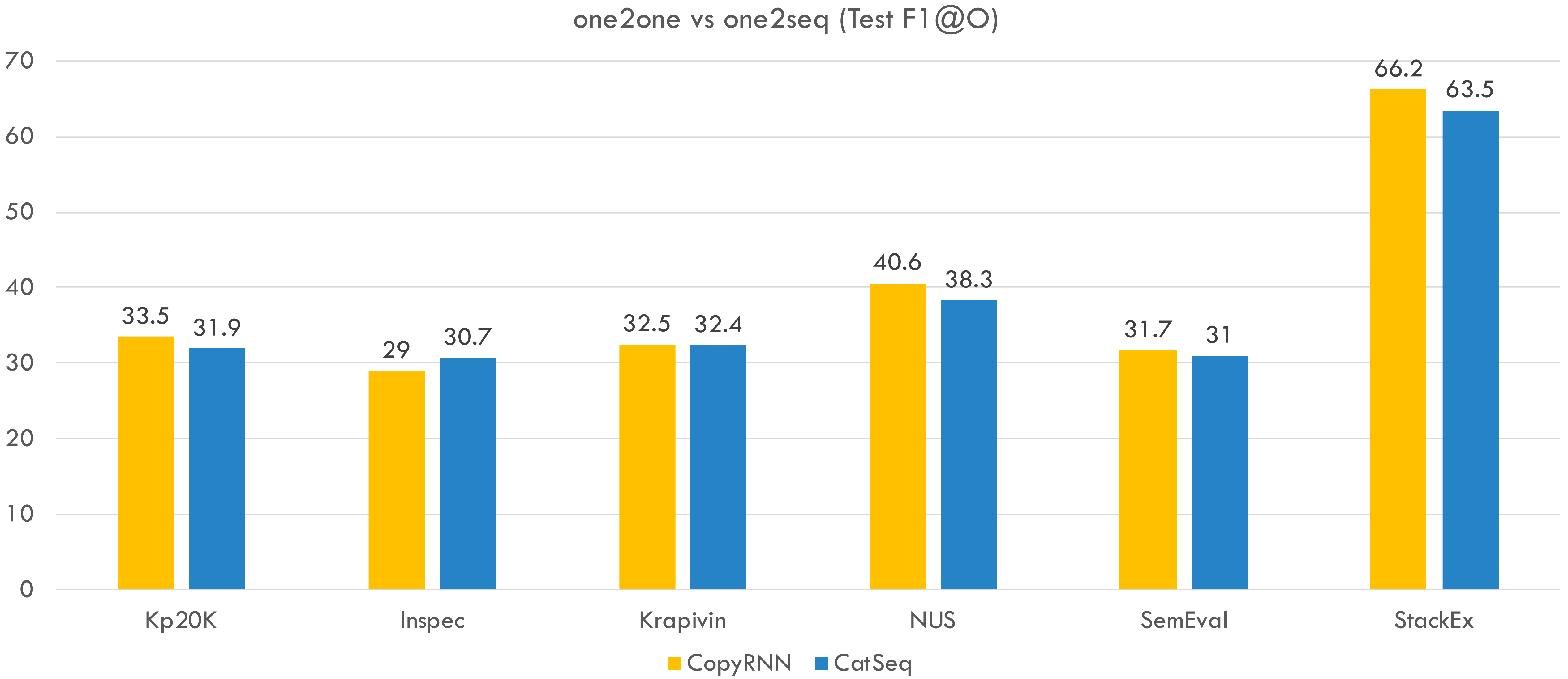
$\mathbf{F}_1@\mathcal{M} \; : \; k = |\hat{\mathcal{Y}}|$

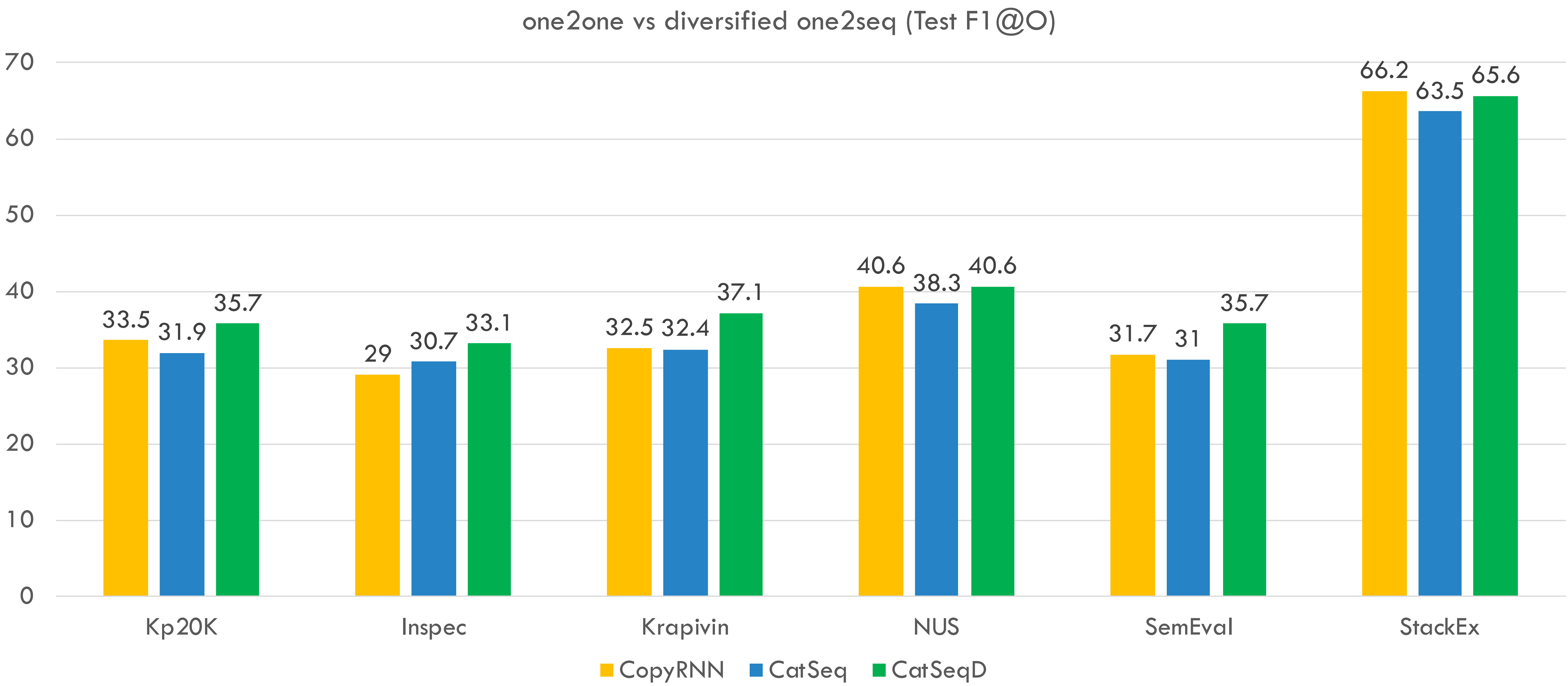
Results and Analyses

Introducing a new keyphrase generation dataset: StackEx

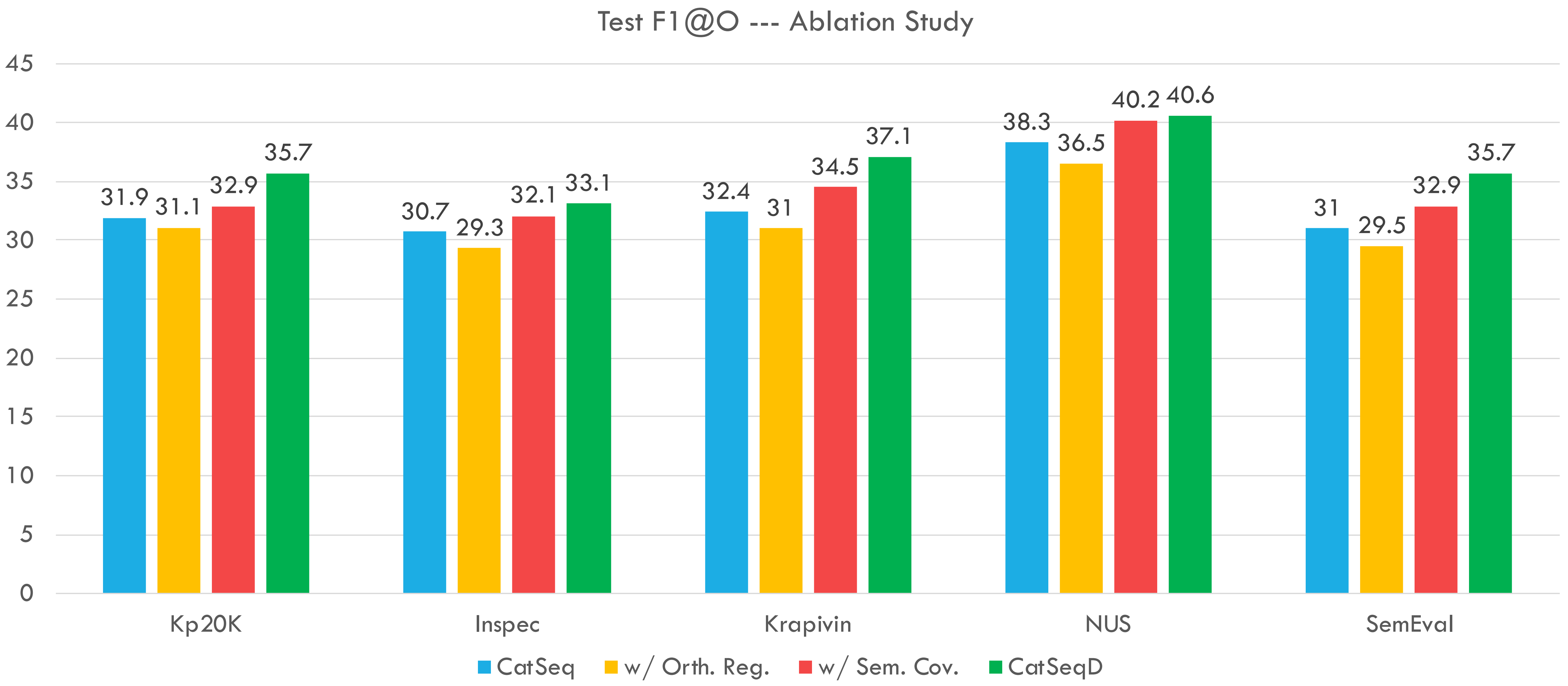
Data Size			Phrase Length		Present/Absent	
#Train	#Valid	#Test	Mean	Var	%Present	% Absent
~298k	~16k	~16k	2.7	1.4	57.5%	42.5%

- As a nice complement to the widely used scientific publication datasets, StackEx is in the domain of community Q&A.
- Due to its unique data collection approach, StackEx has more “common” absent keyphrases.


one2one vs one2seq (Test F1@O)
70
60
50
40
30
20
10
0
33.5
31.9
29
30.7
32.5
32.4
40.6
38.3
31.7
31
66.2
63.5
Kp20K
Inspec
Krapivin
NUS
SemEval
StackEx
CopyRNN
CatSeq

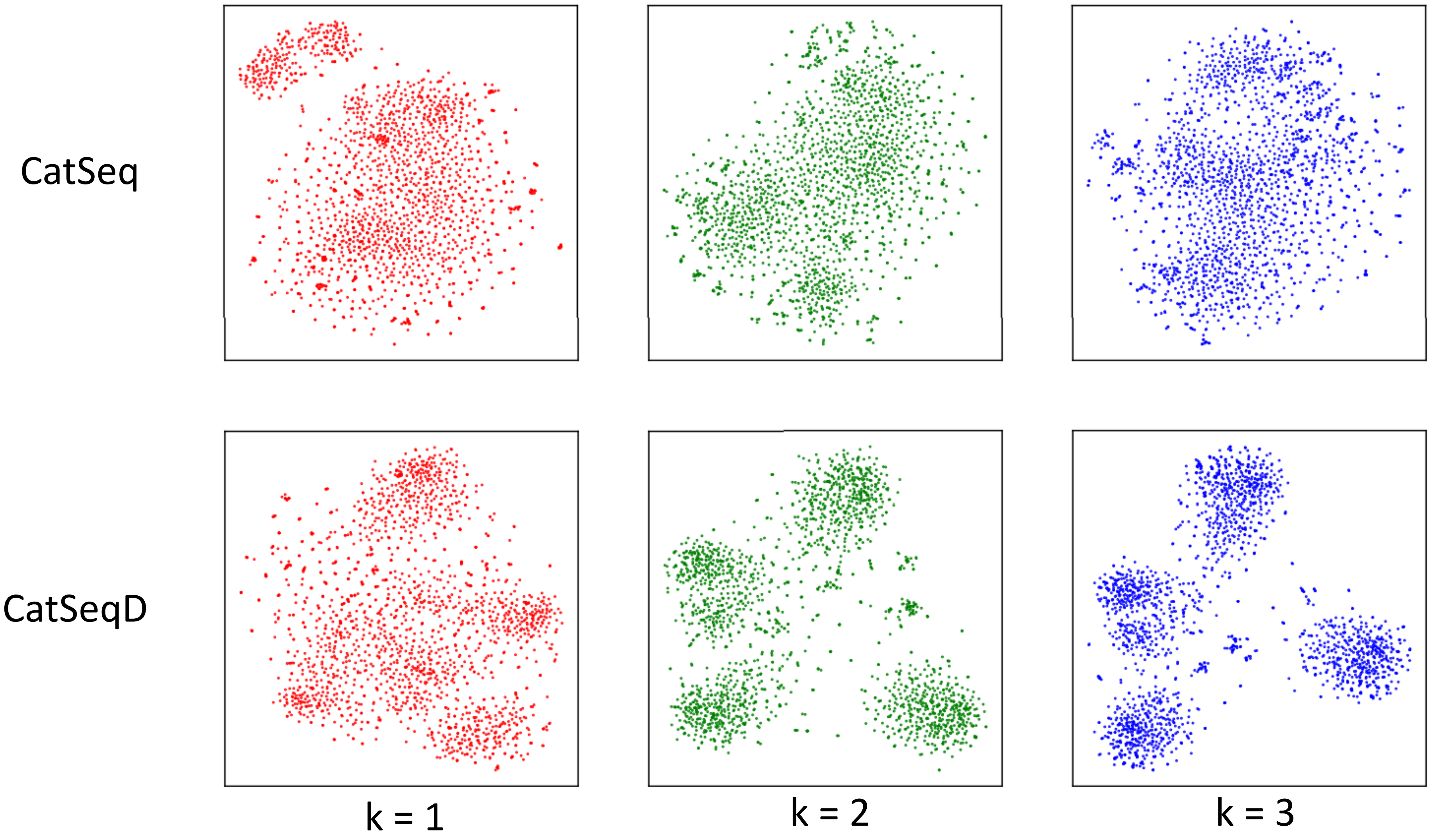


one2one vs diversified one2seq (Test F1@O)
70
60
50
40
30
20
10
0
33.5
31.9
35.7
29
30.7
33.1
32.5
32.4
37.1
40.6
38.3
40.6
31.7
31
35.7
66.2
63.5
65.6
Kp20K
Inspec
Krapivin
NUS
SemEval
StackEx
CopyRNN
CatSeq
CatSeqD



Test F1@O --- Ablation Study
45
40
35
30
25
20
15
10
5
0
31.9
31.1
32.9
35.7
30.7
29.3
32.1
33.1
32.4
31
34.5
37.1
38.3
36.5
40.2
40.6
31
29.5
32.9
35.7
Kp20K
Inspec
Krapivin
NUS
SemEval
CatSeq
w/ Orth. Reg.
w/ Sem. Cov.
CatSeqD

Decoder Hidden States Visualization



k: number of steps after the previous delimiter token

Thank you!

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

QA: Wednesday July 8, 2020. 14A & 15A

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