Keyphrase Generation Based on Deep Seq2seq Model

IEEE Access, VOL. 6, AUG. 2018,
ISBN: 2169-3536
Page(s): 46047 - 46057

Authors: YONG ZHANG, WEIDONG XIAO.

Presenter: WENWEI KANG

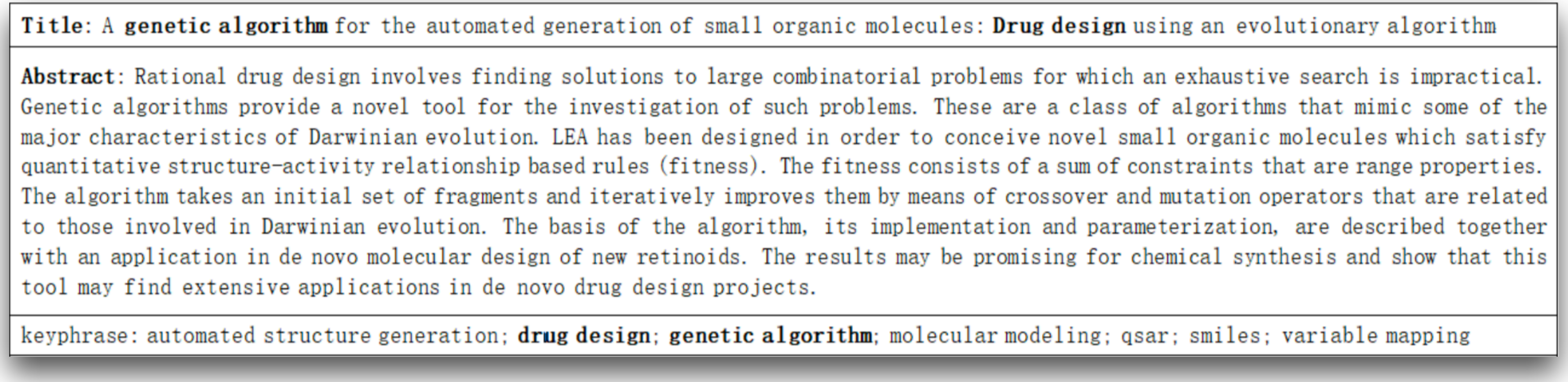
OUTLINE

- Introduction
- Seq2seq
- Mechanisms design
- Evaluation
- Conclusion

Introduction(1/2)

- **Keyphrase**
  - Short texts summarize the information of a document.

- **Applications**
  - Knowledge mining
  - Summarization



Title: A **genetic algorithm** for the automated generation of small organic molecules: **Drug design** using an evolutionary algorithm

Abstract: Rational drug design involves finding solutions to large combinatorial problems for which an exhaustive search is impractical. Genetic algorithms provide a novel tool for the investigation of such problems. These are a class of algorithms that mimic some of the major characteristics of Darwinian evolution. LEA has been designed in order to conceive novel small organic molecules which satisfy quantitative structure-activity relationship based rules (fitness). The fitness consists of a sum of constraints that are range properties. The algorithm takes an initial set of fragments and iteratively improves them by means of crossover and mutation operators that are related to those involved in Darwinian evolution. The basis of the algorithm, its implementation and parameterization, are described together with an application in de novo molecular design of new retinoids. The results may be promising for chemical synthesis and show that this tool may find extensive applications in de novo drug design projects.

keyphrase: automated structure generation; **drug design**; **genetic algorithm**; molecular modeling; qsar; smiles; variable mapping

https://ieeexplore.ieee.org/document/8438457

Introduction(2/2)

- **Previous researches**
  - Ranking and selecting meaningful words from the source text.
  - Be *unable* to generate keyphrases which do not appear in the source text.

- **Recent researches**
  - Generating keyphrases by neural network(seq2seq).
  - Be *able* to generate keypharses which do not appear in the source text.

Seq2seq

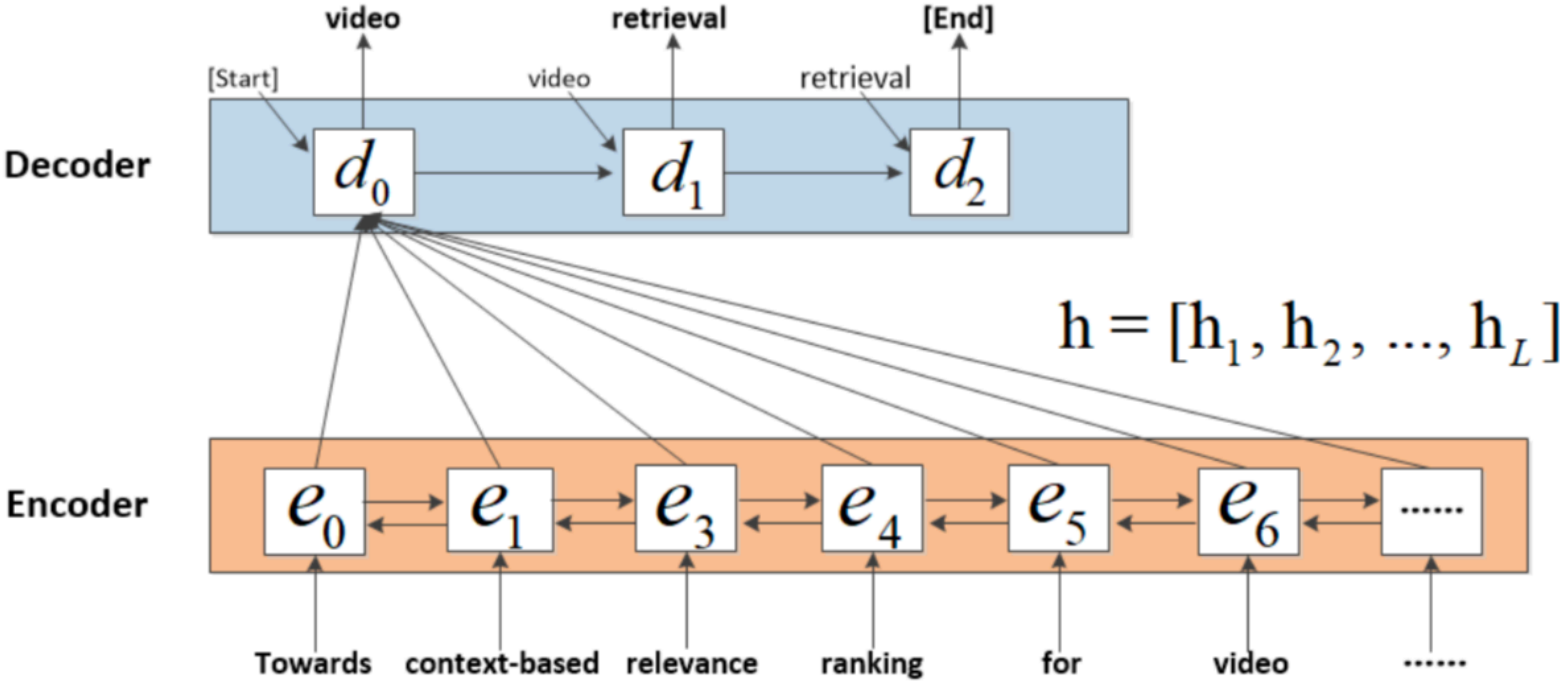
Encoder - decoder model(Seq2seq)

1. Encoder summarizes semantic and generates hidden representations.
2. Decoder gets hidden representations and generates keyphrases.

$$\text{Hidden representations} : \mathrm{h} = [\mathrm{h}_1, \mathrm{h}_2, \ldots, \mathrm{h}_L]$$

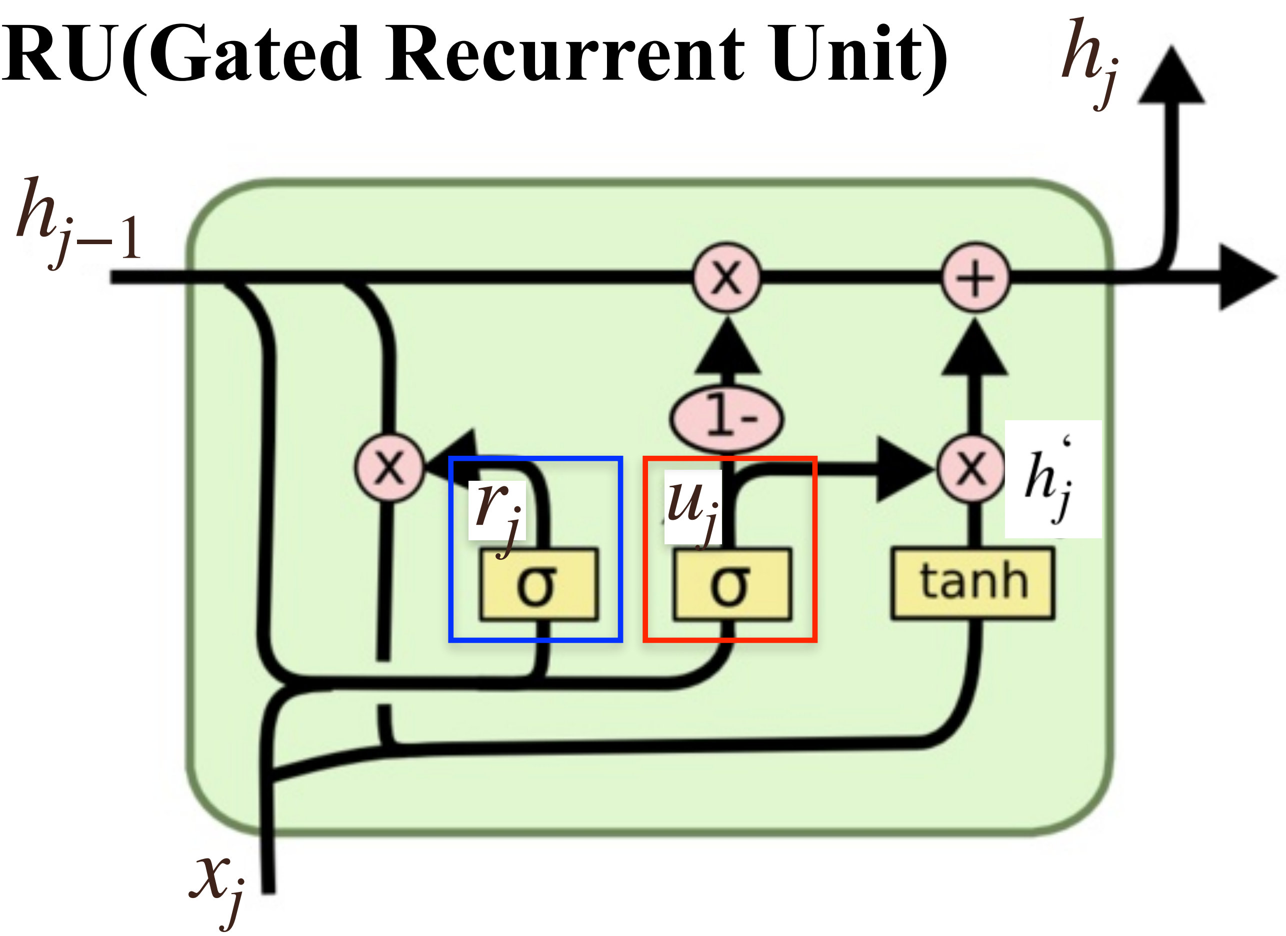
$$\text{Encoder hidden states} : \mathrm{e} = [\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_L]$$

$$\text{Decoder hidden states} : \mathrm{d} = [\mathrm{d}_1, \mathrm{d}_2, \ldots, \mathrm{d}_m]$$



Mechanisms design(1/7)

GRU(Gated Recurrent Unit)



Update gate $\boxed{u_j} = \sigma(W_{ux}x_j + W_{uh}h_{j-1} + b_u)$

Reset gate $\boxed{r_j} = \sigma(W_{rx}x_j + W_{rh}h_{j-1} + b_r)$

State candidate $h_j^{'} = tanh(W_{hx}x_j + W_{hh}(r_j \bigodot h_{j-1} + b_h))$

Current state $h_j = (1 - u_j) \bigodot h_j^{'} + u_j \bigodot h_{j-1}$

$\bigodot$: (Hadamard product), element-wise multiplication

Mechanisms design(2/7)

1. Encoder

Forward: $h_j^f = f(x_j, h_{j-1}^f)$ $\{x_1, x_2, \ldots, x_{\mathrm{Tx}-1}, x_{\mathrm{Tx}}\}$

Backward: $h_j^b = f(x_j, h_{j-1}^b)$ $\{x_{\mathrm{Tx}}, x_{\mathrm{Tx}-1}, \ldots, x_2, x_1\}$

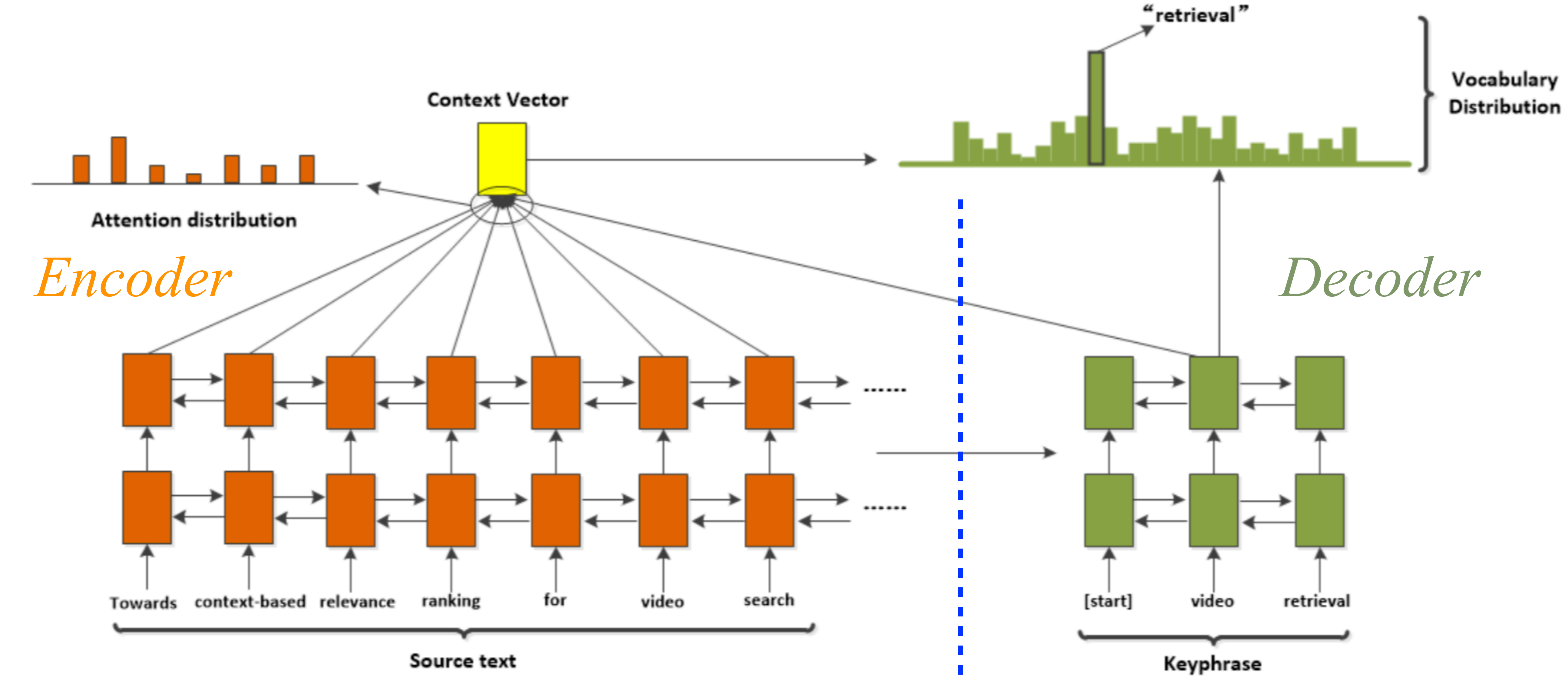
$$h_j = [h_j^f, h_j^b]$$

2. Decoder

Forward: $s_i^f = f(y_i, s_{i-1}^f)$

Backward: $s_i^b = f(y_i, s_{i-1}^b)$

$$s_i = [s_i^f, s_i^b]$$



Mechanisms design(3/7)

3. Attention mechanism

When decoder generates one keyphrase, it considers each time step of encoder.

$$\begin{aligned} e_j^t &= v^T tanh(W_h h_i + W_s s_t + b_{attn}) \\ a^t &= softmax(e^t) \end{aligned}$$

$a^t \in \mathbb{R}^{\text{source text}}$

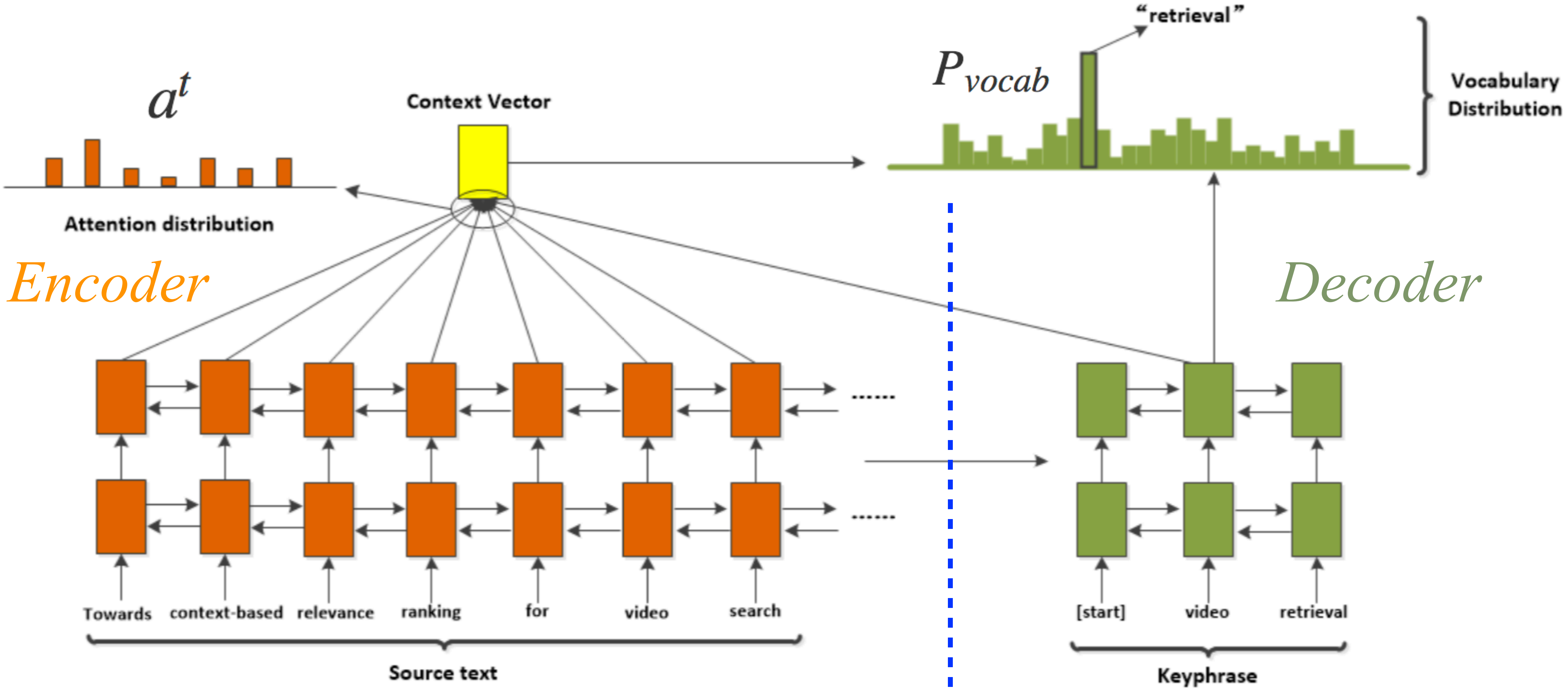
$$\rightarrow h_t^* = \sum_i a_i^t h_i \rightarrow$$

$$P_{vocab} = softmax(V_2 tanh(V_1[s_t, h_t^*] + b_1) + b_2)$$

$P_{vocab} \in \mathbb{R}^{\text{vocab size}}$

$W_h, W_s, b_{attn}, V_1, V_2, b_1, b_2$ is learnable Parameters

V_2 is a matrix to change the dimension to vocabulary size.

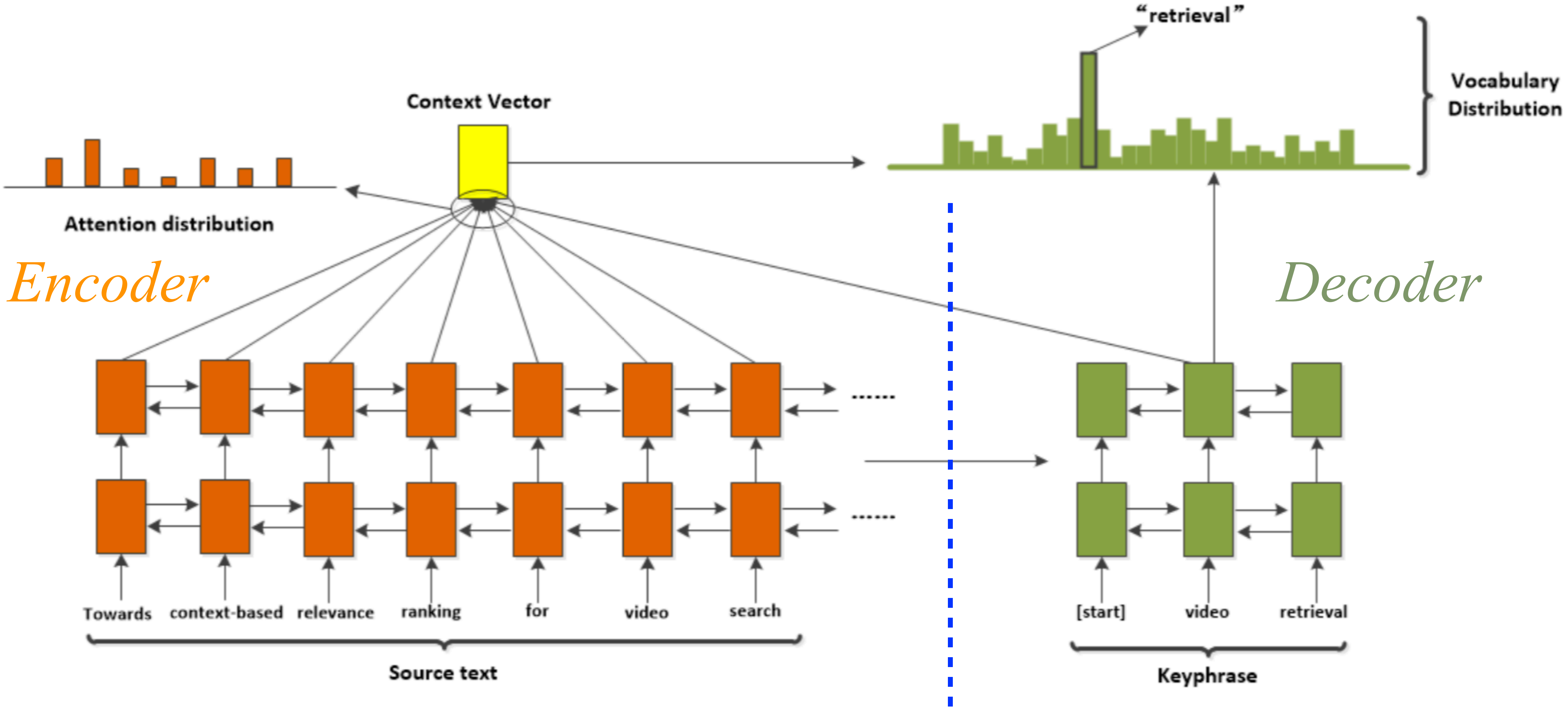


Mechanisms design(4/7)

4. Compute loss

Use P_{vocab} as output: $P(w) = P_{vocab}$

Cross entropy: $loss = -\frac{1}{T}\sum_{t=0}^{T} logP(w_t^*)$



Mechanisms design(5/7)

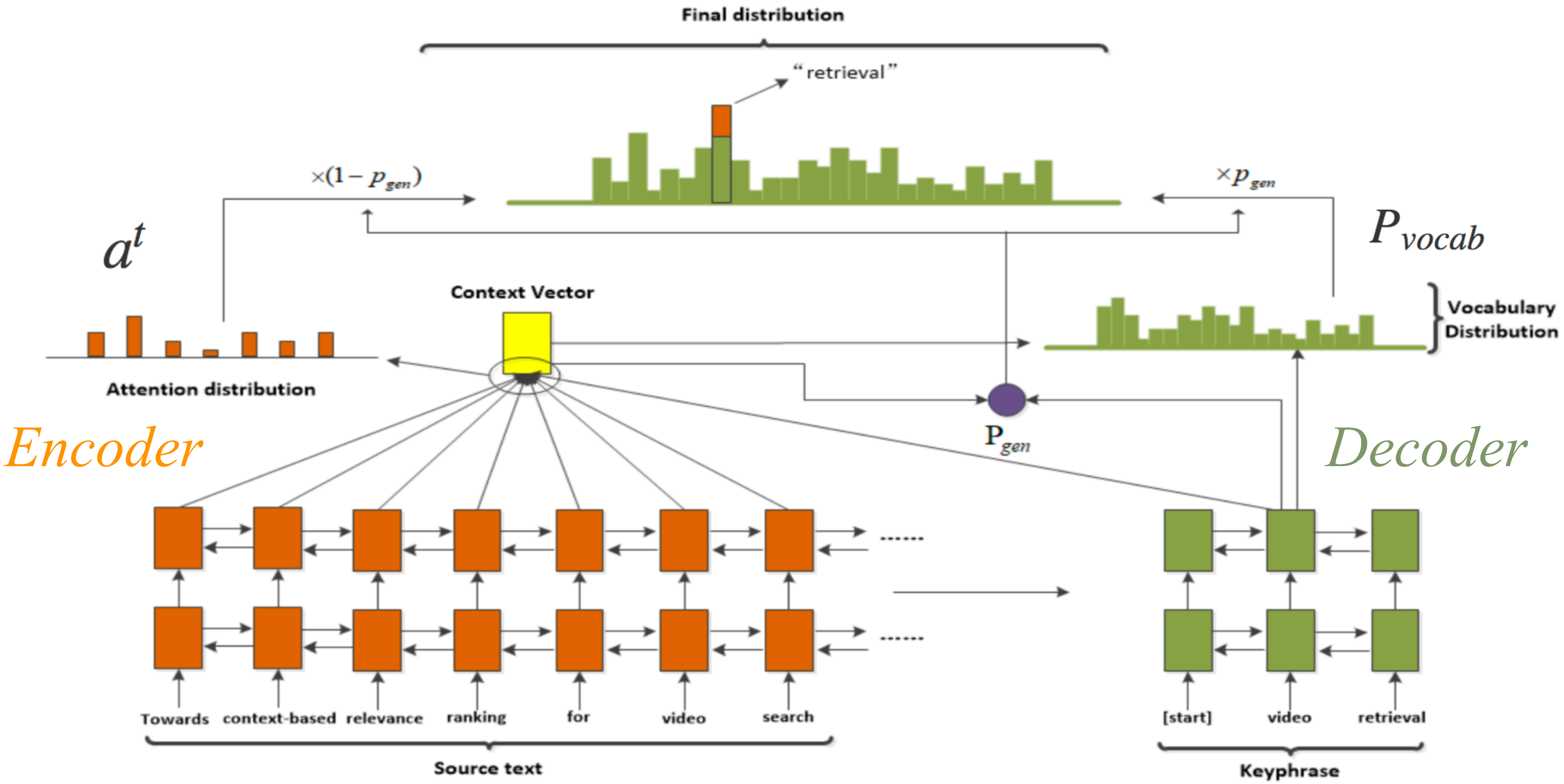
Copy mechanism

Consider the attention and vocabulary distribution together at time t.

$$P(w) = p_{gen} P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} a_i^t$$

$$p_{gen} = \sigma(w_{ch^*} h_t^* + w_{cs} s_t + w_{cy} y_t + b_{gen})$$

$w_{ch}, w_{cs}, w_{cy}, b_{gen}$ is learnable paramters, σ is the sigmoid function.



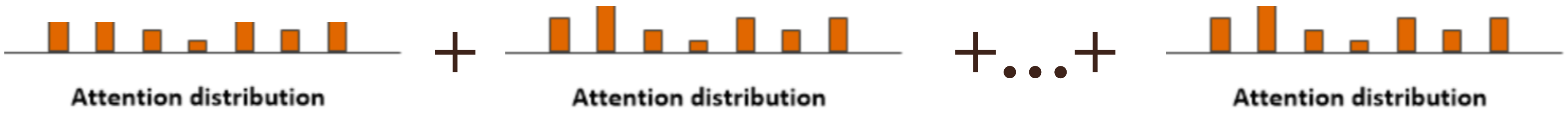
Mechanisms design(6/7)

Coverage mechanism

Decoder will reference the attentions from previous decoder steps and attention distribution.

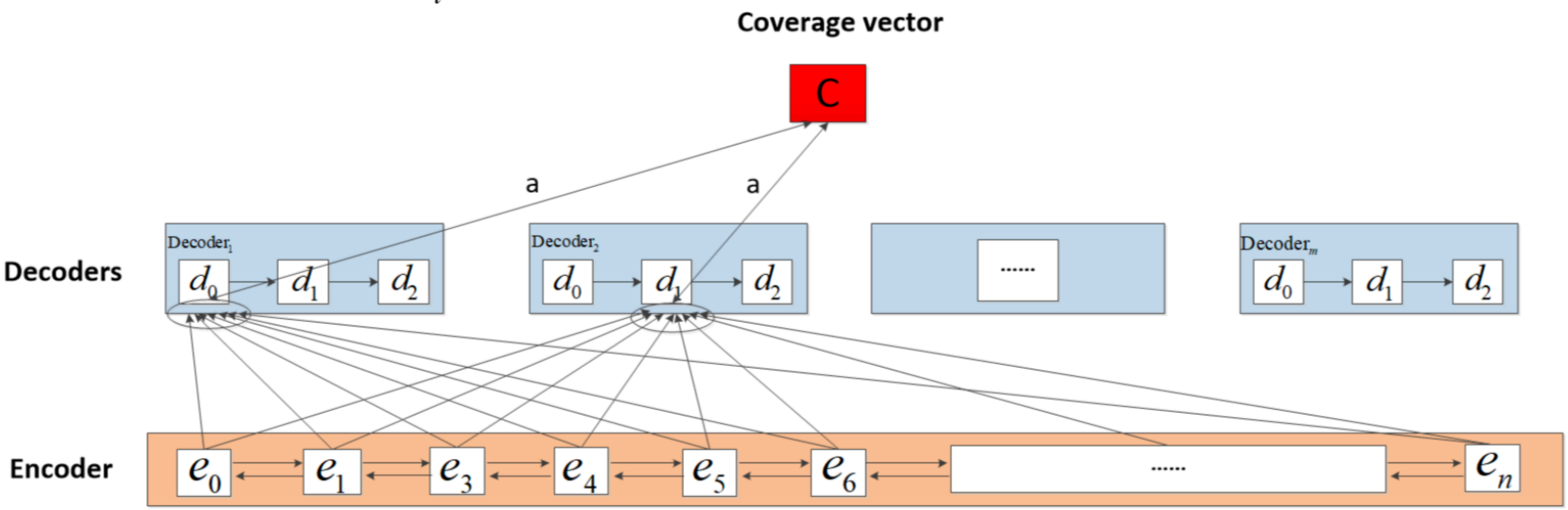
1. Coverage vector: summation over all the attention distribution.

$$c^t = \sum_{t_s=0}^{t-1} a^{t_s}$$



2. Use coverage vector as input.

$$e_j^t = v^T tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{attn})$$



Mechanisms design(7/7)

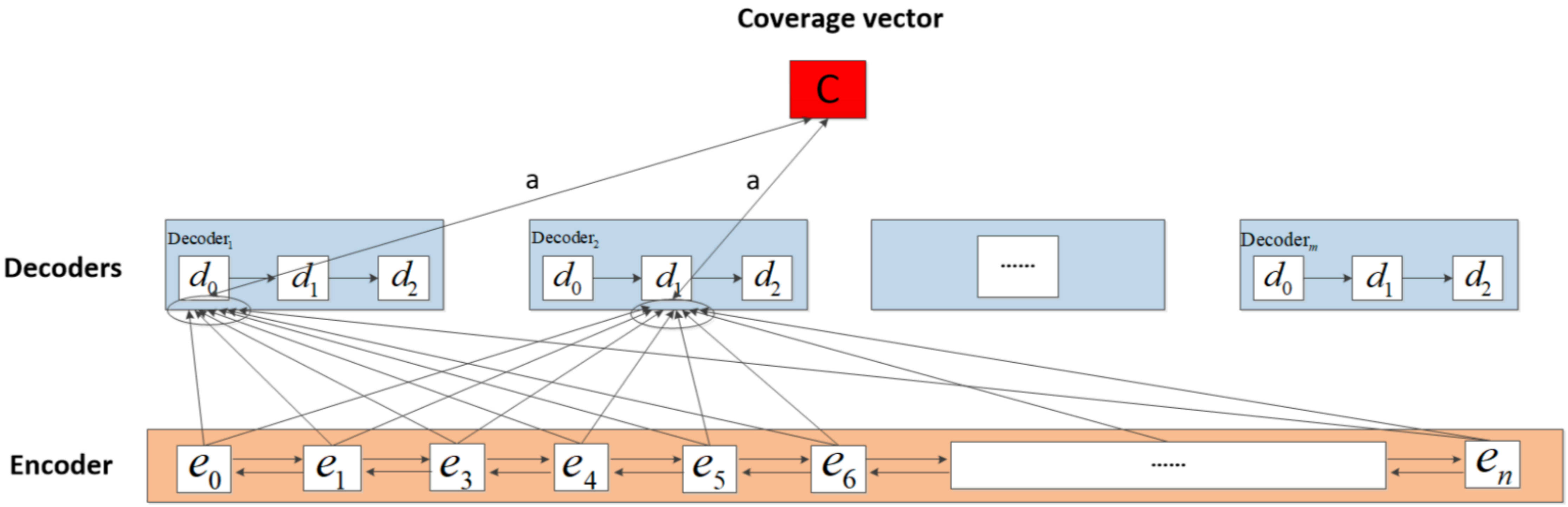
Coverage mechanism

An extra loss function penalizes repeatedly attending to the **same locations.**

$$loss_t = -logP(w_t^*) + \boxed{\lambda \sum_i min(a_i^t, c_i^t)} \quad \Rightarrow \quad loss = \frac{1}{T}\sum_{t=0}^{T} loss_t$$

a_i^t : attention distribution

c_i^t : coverage vector



Evaluation(1/4)

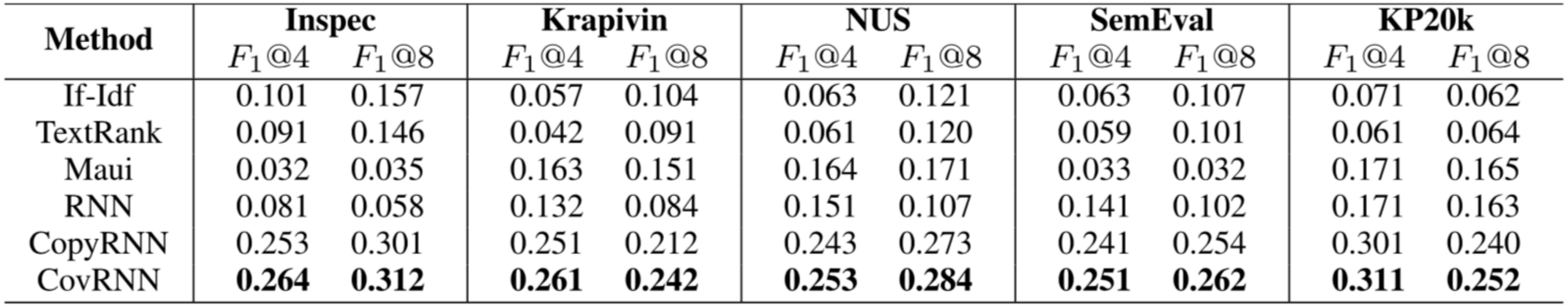
Training set: 527,830 articles
Validation set: 20,000 articles

Testing set:

- KP20k: 20,000 articles with titles, abstracts and keyphrases
- Inspec: 500 paper abstracts
- Krapivin: 400 papers with text and keyphrases
- NUS: 211 papers
- SemEval-2010: 100 articles

Evaluation(2/4)

The performance of predicting keyphrases.

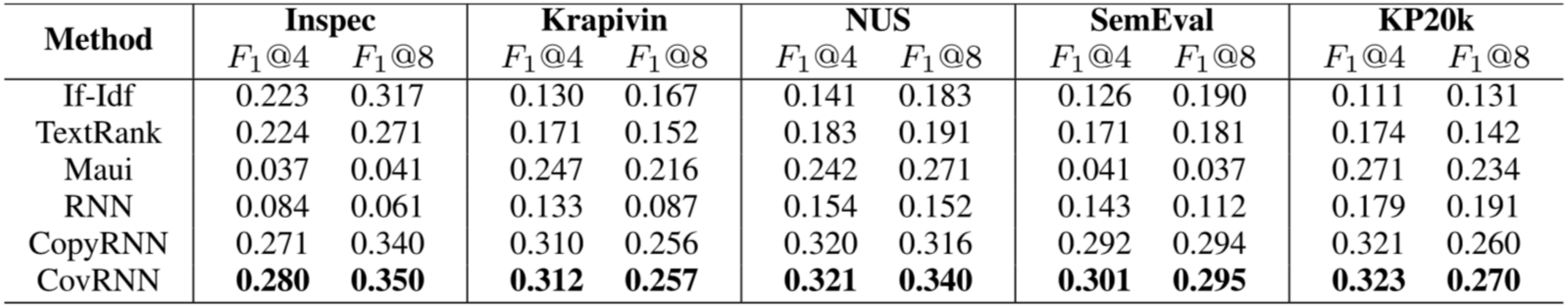


Method	Inspec		Krapivin		NUS		SemEval		KP20k	
	F_1@4	F_1@8	F_1@4	F_1@8	F_1@4	F_1@8	F_1@4	F_1@8	F_1@4	F_1@8
If-Idf	0.101	0.157	0.057	0.104	0.063	0.121	0.063	0.107	0.071	0.062
TextRank	0.091	0.146	0.042	0.091	0.061	0.120	0.059	0.101	0.061	0.064
Maui	0.032	0.035	0.163	0.151	0.164	0.171	0.033	0.032	0.171	0.165
RNN	0.081	0.058	0.132	0.084	0.151	0.107	0.141	0.102	0.171	0.163
CopyRNN	0.253	0.301	0.251	0.212	0.243	0.273	0.241	0.254	0.301	0.240
CovRNN	**0.264**	**0.312**	**0.261**	**0.242**	**0.253**	**0.284**	**0.251**	**0.262**	**0.311**	**0.252**

- RNN: Seq2seq with attention mechanism
- CopyRNN: Seq2seq with attention mechanism, copy mechanism
- CovRNN: Seq2seq with attention mechanism, copy mechanism, coverage mechanism

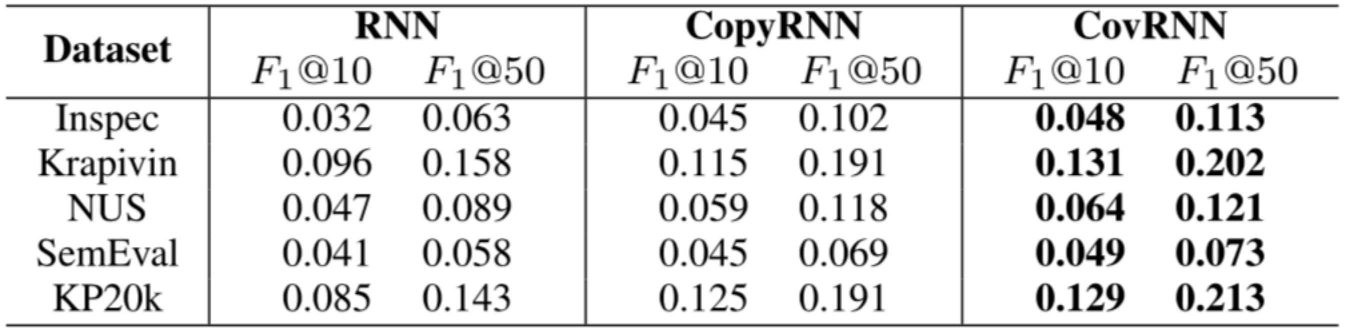
Evaluation(3/4)

The performance of predicting present keyphrases.



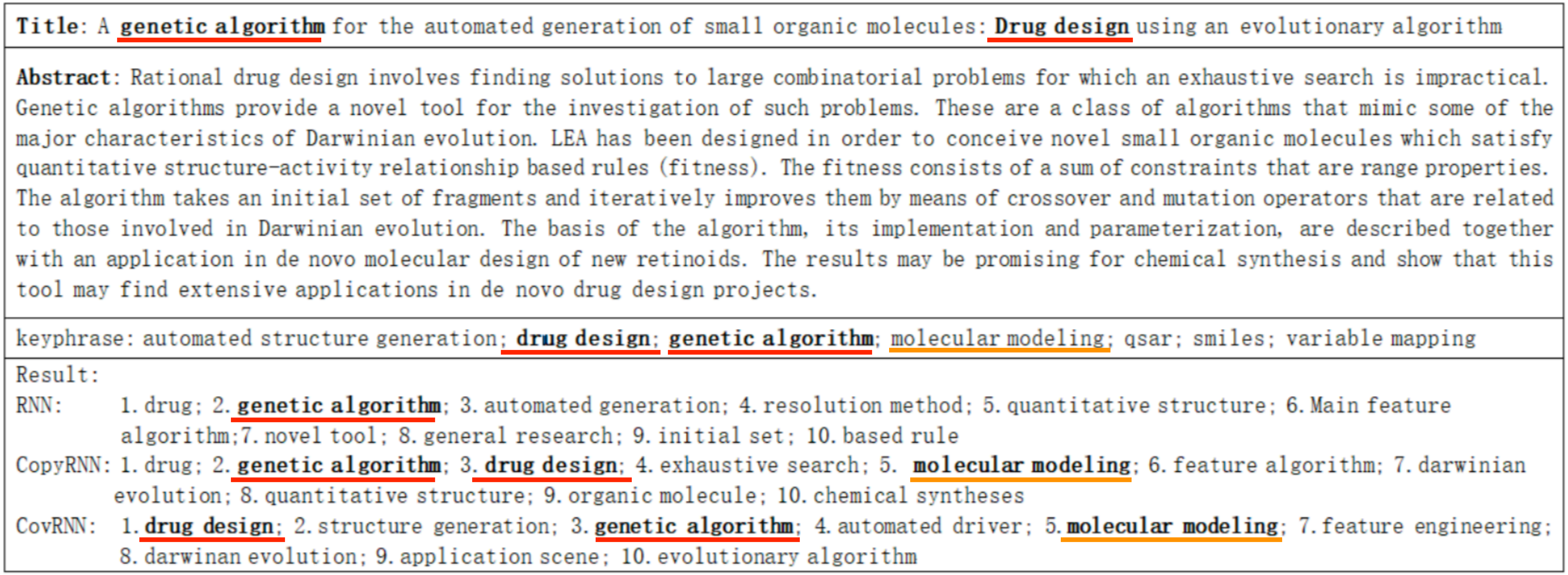
Method	Inspec		Krapivin		NUS		SemEval		KP20k	
	F_1@4	F_1@8	F_1@4	F_1@8	F_1@4	F_1@8	F_1@4	F_1@8	F_1@4	F_1@8
If-Idf	0.223	0.317	0.130	0.167	0.141	0.183	0.126	0.190	0.111	0.131
TextRank	0.224	0.271	0.171	0.152	0.183	0.191	0.171	0.181	0.174	0.142
Maui	0.037	0.041	0.247	0.216	0.242	0.271	0.041	0.037	0.271	0.234
RNN	0.084	0.061	0.133	0.087	0.154	0.152	0.143	0.112	0.179	0.191
CopyRNN	0.271	0.340	0.310	0.256	0.320	0.316	0.292	0.294	0.321	0.260
CovRNN	**0.280**	**0.350**	**0.312**	**0.257**	**0.321**	**0.340**	**0.301**	**0.295**	**0.323**	**0.270**

The performance of predicting absent keyphrases.



Dataset	RNN		CopyRNN		CovRNN	
	F_1@10	F_1@50	F_1@10	F_1@50	F_1@10	F_1@50
Inspec	0.032	0.063	0.045	0.102	**0.048**	**0.113**
Krapivin	0.096	0.158	0.115	0.191	**0.131**	**0.202**
NUS	0.047	0.089	0.059	0.118	**0.064**	**0.121**
SemEval	0.041	0.058	0.045	0.069	**0.049**	**0.073**
KP20k	0.085	0.143	0.125	0.191	**0.129**	**0.213**

Evaluation(4/4)



Title: A **genetic algorithm** for the automated generation of small organic molecules: **Drug design** using an evolutionary algorithm

Abstract: Rational drug design involves finding solutions to large combinatorial problems for which an exhaustive search is impractical. Genetic algorithms provide a novel tool for the investigation of such problems. These are a class of algorithms that mimic some of the major characteristics of Darwinian evolution. LEA has been designed in order to conceive novel small organic molecules which satisfy quantitative structure-activity relationship based rules (fitness). The fitness consists of a sum of constraints that are range properties. The algorithm takes an initial set of fragments and iteratively improves them by means of crossover and mutation operators that are related to those involved in Darwinian evolution. The basis of the algorithm, its implementation and parameterization, are described together with an application in de novo molecular design of new retinoids. The results may be promising for chemical synthesis and show that this tool may find extensive applications in de novo drug design projects.

keyphrase: automated structure generation; **drug design**; **genetic algorithm**; molecular modeling; qsar; smiles; variable mapping

Result:

RNN: 1. drug; 2. **genetic algorithm**; 3. automated generation; 4. resolution method; 5. quantitative structure; 6. Main feature algorithm;7. novel tool; 8. general research; 9. initial set; 10. based rule

CopyRNN: 1. drug; 2. **genetic algorithm**; 3. **drug design**; 4. exhaustive search; 5. **molecular modeling**; 6. feature algorithm; 7. darwinian evolution; 8. quantitative structure; 9. organic molecule; 10. chemical syntheses

CovRNN: 1. **drug design**; 2. structure generation; 3. **genetic algorithm**; 4. automated driver; 5. **molecular modeling**; 7. feature engineering; 8. darwinan evolution; 9. application scene; 10. evolutionary algorithm

Top10 predictions for above three seq2seq models.

Keyphrase in bold: keyphrases appear in the source text.

Conclusion

- The proposed model with attention, copy and coverage mechanism is able to deal with OOV(out of vocabulary) problem and reduce information redundancy between different keyphrases.

- How to determine top-n keyphrases completely express the meaning of text still is a problem.