Text summarization

Yufeng Lv

10/15/2018

Types

No.	Types of summary	Factors
1.	Single and multi-document	Number of documents
2.	Extractive and abstractive	Output (if extract or abstract is required)
3.	Generic and query-focused	Purpose (whether general or query related data is required)
4.	Supervised and unsupervised	Availability of training data
5.	Mono, multi and cross-lingual	Language
6.	Web-based	For summarizing web pages
7.	E-mail based	For summarizing e-mails
8.	Personalized	Information specific to a user's need
9.	Update	Current updates regarding a topic
10.	Sentiment-based	Opinions are detected
11.	Survey	Important facts regarding person, place or any other entity

Types

No.	Types of summary	Factors
1.	Single and multi-document	Number of documents
2.	Extractive and abstractive	Output (if extract or abstract is required)
3.	Generic and query-focused	Purpose (whether general or query related data is required)
4.	Supervised and unsupervised	Availability of training data
5.	Mono, multi and cross-lingual	Language
6.	Web-based	For summarizing web pages
7.	E-mail based	For summarizing e-mails
8.	Personalized	Information specific to a user's need
9.	Update	Current updates regarding a topic
10.	Sentiment-based	Opinions are detected
11.	Survey	Important facts regarding person, place or any other entity

Data sets

- DUC(2001-2007)/TAC(2008-)
  - Text summarization competition dataset, for model evaluation
- Gigaword
  - include 9,500,000 news articles, using headline for summary
  - for single sentence summarization
- CNN/Daily Mail
  - multi-sentence summary
- Large Scale Chinese Short Text summarization Dataset(LCSTS)
  - A Short Text Chinese dataset, from Sina Weibo.
  - 2,400,591 short text, summary pairs/10,666 human labeled pairs

Hu B, Chen Q, Zhu F. Lcsts: A large scale chinese short text summarization dataset[J]. arXiv preprint arXiv:1506.05865, 2015.

【俄罗斯申请加入联合国人权理事会】美国宣布退出联合国人权理事会不久后，俄罗斯常驻联合国代表团于当地时间周三（20日）表示，俄罗斯已经申请成为联合国人权理事会2021-2023届成员国。据俄罗斯卫星通讯社报道，俄罗斯常驻联合国代表团第一秘书Fedor Strzhizhovskiy表示，俄罗斯想要继续在人权理事会 ... 全文

val.txt.tgt.tagged — 已编辑

```
<t> a man in suburban boston is selling snow online to customers in warmer states . </t> <t> for $ 89 , he will ship 6 pounds of snow in an insulated styrofoam box . </t>
```

val.txt.src — 已编辑

```
-lrb- cnn -rrb- the only thing crazier than a guy in snowbound massachusetts boxing up the powdery white stuff and offering it for sale online ? people are actually buying it . for $ 89 , self-styled entrepreneur kyle waring will ship you 6 pounds of boston-area snow in an insulated styrofoam box -- enough for 10 to 15 snowballs , he says . but not if you live in new england or surrounding states . `` we will not ship snow to any states in the northeast ! '' says waring 's website , shipsnowyo.com . `` we 're in the business of expunging snow ! '' his website and social media accounts claim to have filled more than 133 orders for snow -- more than 30 on tuesday alone , his busiest day yet . with more than 45 total inches , boston has set a record this winter for the snowiest month in its history . most residents see the huge piles of snow choking their yards and sidewalks as a nuisance , but waring saw an opportunity . according to boston.com , it all started a few weeks ago , when waring and his wife were shoveling deep snow from their yard in manchester-by-the-sea , a coastal suburb north of boston . he joked about shipping the stuff to friends and family in warmer states , and an idea was born . his business slogan : `` our nightmare is your dream ! '' at first , shipsnowyo sold snow packed into empty 16.9-ounce water bottles for $ 19.99 , but the snow usually melted before it reached its destination . so this week , waring began shipping larger amounts in the styrofoam cubes , which he promises will arrive anywhere in the u.s. in less than 20 hours . he also has begun selling a 10-pound box of snow for $ 119 . many of his customers appear to be companies in warm-weather states who are buying the snow as a gag , he said . whether waring can sustain his gimmicky venture into the spring remains to be seen . but he has no shortage of product . `` at this rate , it 's going to be july until the snow melts , '' he told boston.com . `` but i 've thought about taking this idea and running with it for other seasonal items . maybe i 'll ship some fall foliage . ''
```

```
<DOC id="XIN_CMN_19980201.0003" type="story">
<HEADLINE>
李岚清会见欧盟委员会主席桑特
</HEADLINE>
<DATELINE>
新华社达沃斯(瑞士)2月1日电
</DATELINE>
<TEXT>
<P>
(记者陈维斌　严
明)正在瑞士达沃斯出席世界经济论坛年会的中国国务院副总理李岚清1日在这里会见了欧盟委员会主席桑特。
</P>
<P>
李岚清说,近年来,中国同欧盟及其成员国的关系继续保持良好的发展势头,双方高层互访和接触频繁,不同层次的政治磋商和对话活跃,各个领域的合作与交流不断扩大。双方经济互补性强,对许多重大国际问题有着一致或相似的看法。
</P>
<P>
桑特说,欧盟非常重视发展对华关系,对中国改革开放取得的重大成就深感钦佩。
</P>
<P>
在谈到中国和欧盟的经济关系时,李岚清表示,不久前欧盟委员会建议不要将中国划归"非市场经济"国家,希望欧盟委员会积极推动欧盟理事会通过这一建议。桑特表示,相信这一问题会很快得到解决。他说,欧盟希望进一步加强同发展中国家之间业已非常密切的经贸合作关系。
</P>
<P>
在谈到中国加入世贸组织的问题时,李岚清表示,希望欧盟以更加灵活、务实和建设性的态度解决谈判中悬而未决的问题,促使谈判早日结束。桑特重申,欧盟支持并希望中国能尽快加入世贸组织。
</P>
</TEXT>
</DOC>
```

Evaluation——ROUGE

- Rouge-N (unigram, bigrams, trigrams, etc)
- Rouge-L (summary level LCS)
- Rouge-W(Weighted LCS)
- Rouge-S (skip-gram)
- Rouge-SU (skip-gram with unigrams)

Flick C. ROUGE: A Package for Automatic Evaluation of summaries[C] The Workshop on Text Summarization Branches Out. 2004:10.

Rouge-N

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

Flick C. ROUGE: A Package for Automatic Evaluation of summaries[C] The Workshop on Text Summarization Branches Out. 2004:10.

Rouge-N

- Auto summary: *the cat was found under the bed*
- Ref summary: *the cat was under the bed*

No	1-gram	ref 1-gram	2-gram	ref 2-gram
1	the	the	the cat	the cat
2	cat	cat	cat was	cat was
3	was	was	was found	was under
4	found	under	found under	under the
5	under	the	under the	the bed
6	the	bed	the bed	
7	bed			
count	7	6	6	5

$$\text{Rouge1}=\frac{6}{6}=1$$

$$\text{Rouge2}=\frac{4}{5}=0.8$$

Flick C. ROUGE: A Package for Automatic Evaluation of summaries[C] The Workshop on Text Summarization Branches Out. 2004:10.

Rouge-L(Longest Common Subsequence)

$$R_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{m} \quad (5)$$

$$P_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{n} \quad (6)$$

$$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \quad (7)$$

Flick C. ROUGE: A Package for Automatic Evaluation of summaries[C] The Workshop on Text Summarization Branches Out. 2004:10.

Rouge-W

X: [A B C D E F G]
Y_1: [A B C D H I K]
Y_2: [A H B K C I D]

$$R_{wlcs} = f^{-1}\left(\frac{WLCS(X,Y)}{f(m)}\right) \quad (13)$$

$$P_{wlcs} = f^{-1}\left(\frac{WLCS(X,Y)}{f(n)}\right) \quad (14)$$

$$F_{wlcs} = \frac{(1+\beta^2)R_{wlcs}P_{wlcs}}{R_{wlcs} + \beta^2 P_{wlcs}} \quad (15)$$

Flick C. ROUGE: A Package for Automatic Evaluation of summaries[C] The Workshop on Text Summarization Branches Out. 2004:10.

Rouge-S/Rouge-SU

$$R_{skip2} = \frac{SKIP2(X,Y)}{C(m,2)} \quad (16)$$

$$P_{skip2} = \frac{SKIP2(X,Y)}{C(n,2)} \quad (17)$$

$$F_{skip2} = \frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2}+\beta^2 P_{skip2}} \quad (18)$$

To achieve this, we extend ROUGE-S with the addition of unigram as counting unit.

Flick C. ROUGE: A Package for Automatic Evaluation of summaries[C] The Workshop on Text Summarization Branches Out. 2004:10.

PyRouge

```
from pyrouge import Rouge155

r = Rouge155()
r.system_dir = 'path/to/system_summaries'
r.model_dir = 'path/to/model_summaries'
r.system_filename_pattern = 'some_name.(\d+).txt'
r.model_filename_pattern = 'some_name.[A-Z].#ID#.txt'

output = r.convert_and_evaluate()
print(output)
output_dict = r.output_to_dict(output)
```

Useful tools

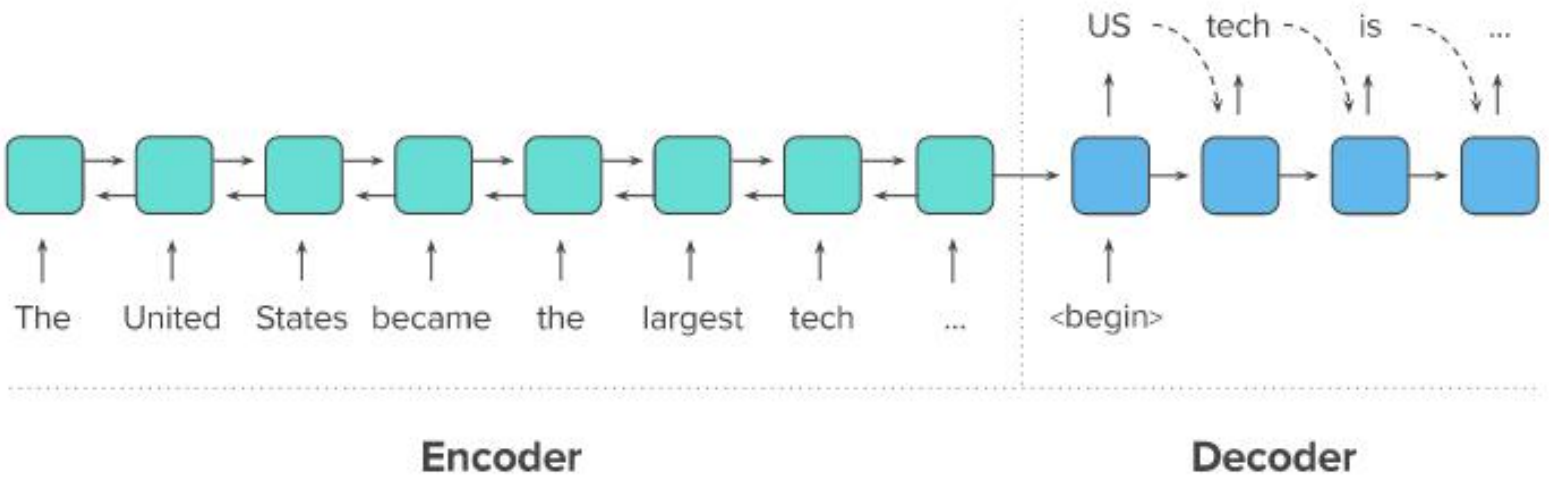
- PKUSUMSUM:
  - Single-document
  - Multi-document
  - Topic-based Multi-document
- HanNLP
  - TextRank

Method
Coverage
Lead
Centroid [1]
TextRank [2]
LexPageRank[3]
ILP [4]
Submodular1 [5]
Submodular2 [6]
ClusterCMRW[7]
ManifoldRank[8]

State of the art

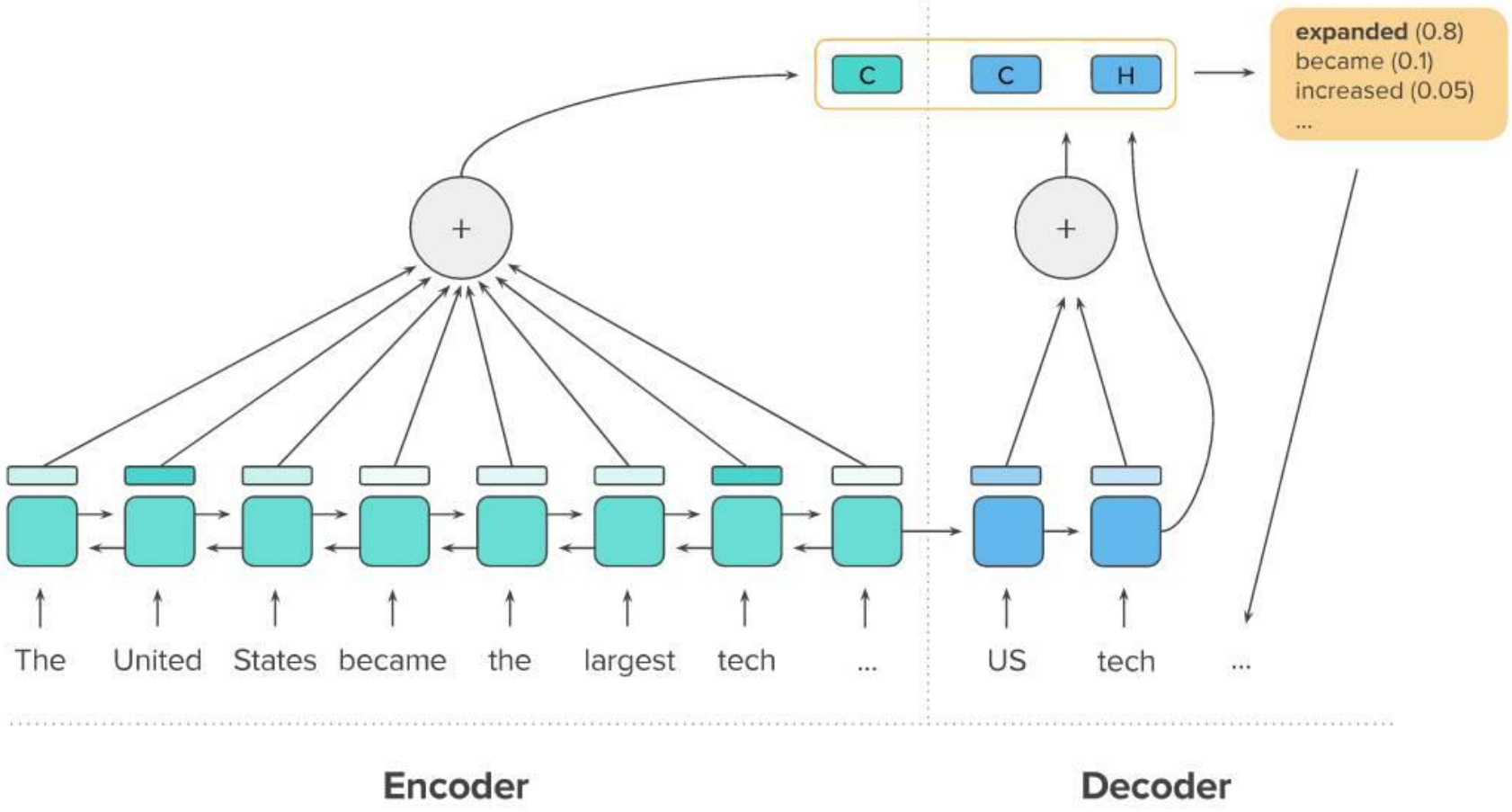
- RNN based
- CNN based
- Attention
- Reinforement

RNN based



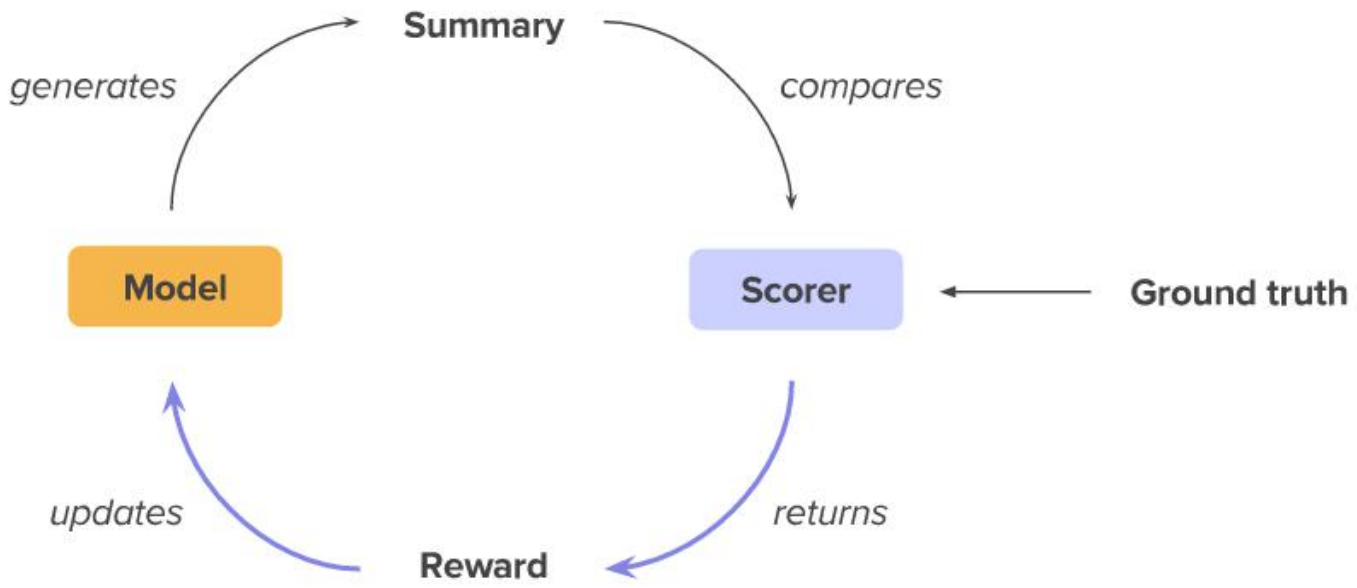
Paulus R, Xiong C, Socher R. A deep reinforced model for abstractive summarization[J]. arXiv preprint arXiv:1705.04304, 2017.

RNN + Attention



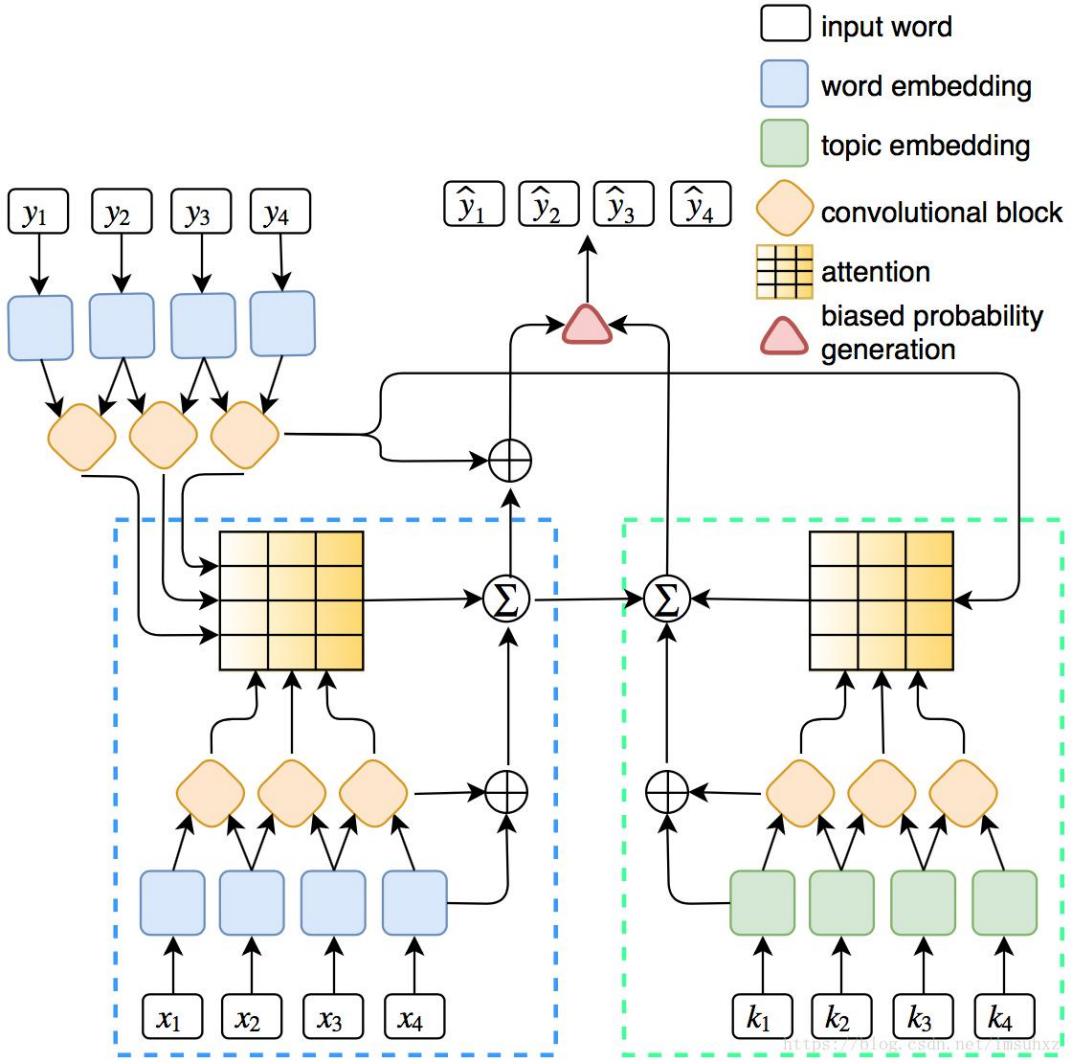
Paulus R, Xiong C, Socher R. A deep reinforced model for abstractive summarization[J]. arXiv preprint arXiv:1705.04304, 2017.

Model + Reinforement learning

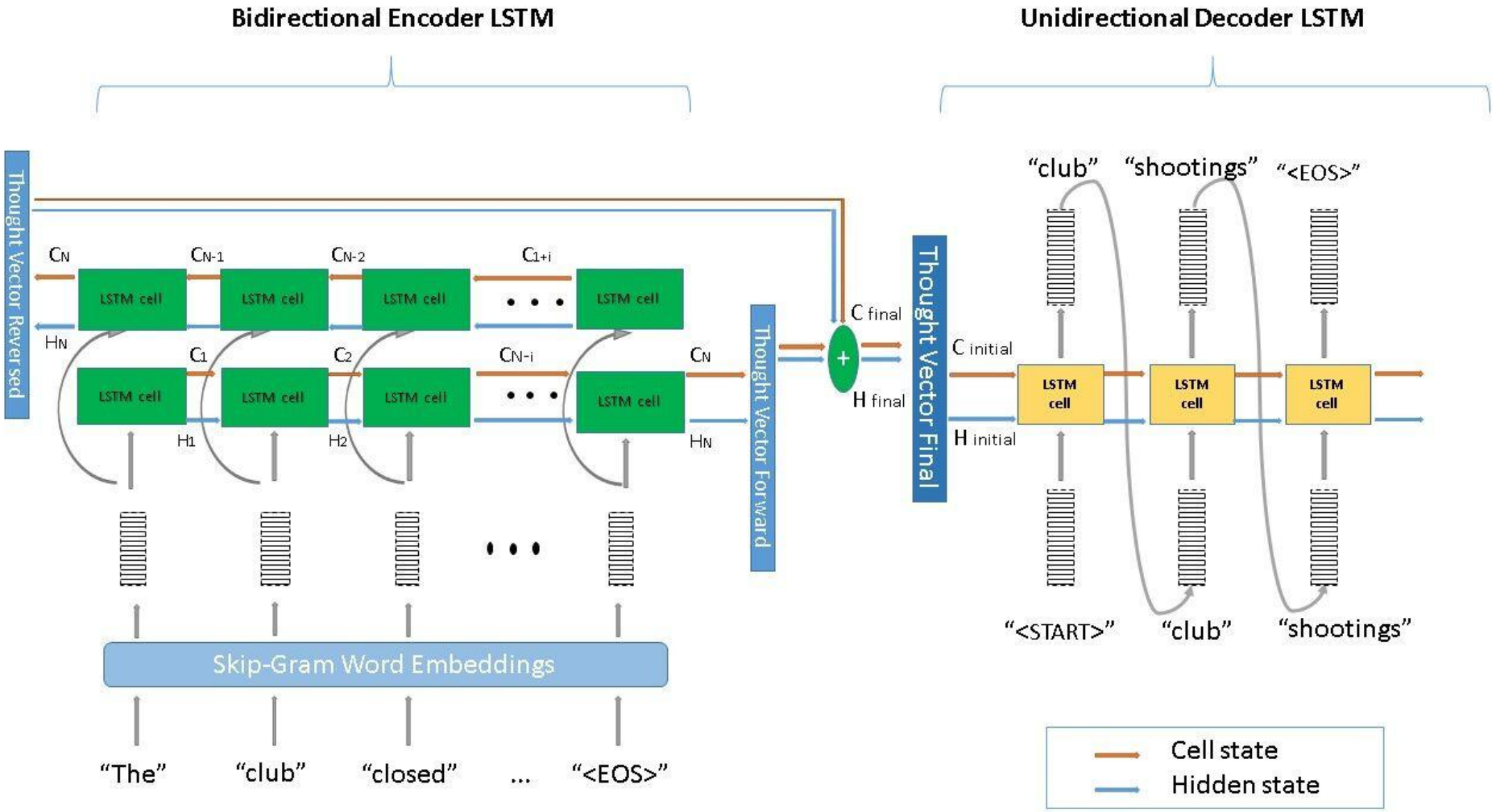


Paulus R, Xiong C, Socher R. A deep reinforced model for abstractive summarization[J]. arXiv preprint arXiv:1705.04304, 2017.

ConvS2S based



Wang L, Yao J, Tao Y, et al. A Reinforced Topic-Aware Convolutional Sequence-to-Sequence Model for Abstractive Text Summarization[J]. 2018.


Bidirectional Encoder LSTM
Unidirectional Decoder LSTM
Thought Vector Reversed
Thought Vector Forward
Thought Vector Final
LSTM cell
CN
CN-1
CN-2
C1+i
C1
C2
CN-i
HN
H1
H2
C final
H final
C initial
H initial
"club"
"shootings"
"<EOS>"
"<START>"
Skip-Gram Word Embeddings
"The"
"club"
"closed"
...
"<EOS>"
Cell state
Hidden state

Our destiny

- Combine UGC and PGC
  - Extractive or abstractive?
  - UGC from Zhihu, Baidu Baike, etc. —— unsupervised
  - UGC + PGC —— multi-document
  - PGC——Long Text

  - The other problem, how to get the structured UGC data for training

Method

- Two baseline:
  - RNN+Attention
  - ConvS2S+Attention
- GANs?
  - Wang Y S, Lee H Y. Learning to Encode Text as Human-Readable Summaries using Generative Adversarial Networks[J]. 2018.
- Joint learning(multi-task)?
  - Salesfore's decaNLP: 10 NLP task for 1 model
- Some awesome ideas for specific issues
  - topic attention
  - Minimum Risk Training

Discussion?