CopyNET - Incorporating Copying Mechanism in Seq2Seq Learning

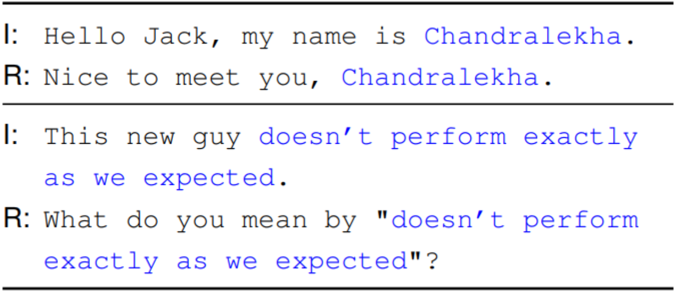
Jiatao Gu, Zhengdong Lu, Hang Li, Victor O.K. Li 2016

Louis (Yiqing) Luo

Aug 3rd, 2018

Motivation

- Problem: Copying in Seq2Seq
  - certain segments in the input sequence are selectively replicated in the output sequence
  - eg. humans tend to repeat ntity names or even long phrases in conversation



I: Hello Jack, my name is Chandralekha.

R: Nice to meet you, Chandralekha.

I: This new guy doesn't perform exactly as we expected.

R: What do you mean by "doesn't perform exactly as we expected"?

Motivation

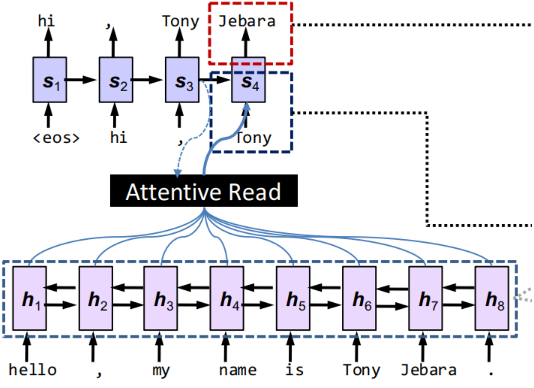
- Problem: Copying in Seq2Seq
  - certain segments in the input sequence are selectively replicated in the output sequence
  - eg. humans tend to repeat entity names or even long phrases in conversation

Proposed Solution:

CopyNET = RNN Encoder & Decoder + Copying Mechanism

RNNSearch: RNN Encoder and Decoder

- typically used in Seq2Seq learning



(a) Attention-based Encoder-Decoder (RNNSearch)

CopyNET

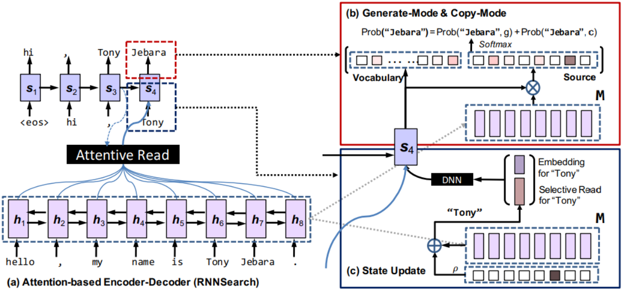
Encoder same. Decoder Differences:

1. **Prediction:** COPYNET predicts words based on a mixed probabilistic model of two modes
2. **State Update:** uses not only its word-embedding but also its corresponding location-specific hidden state in M

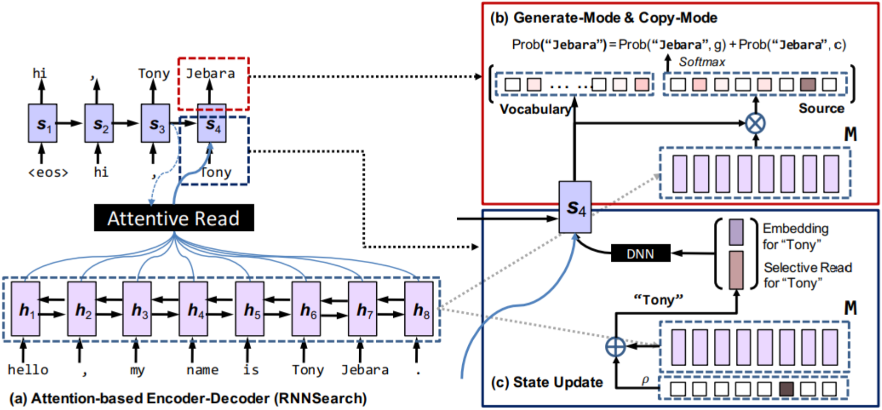
CopyNET

Encoder same. Decoder Differences:

1. **Prediction:** COPYNET predicts words based on a mixed probabilistic model of two modes
2. **State Update:** uses not only its word-embedding but also its corresponding location-specific hidden state in M



Model



(b) Generate-Mode & Copy-Mode
Prob("Jebara") = Prob("Jebara", g) + Prob("Jebara", c)
Softmax
Vocabulary
Source
M
hi
,
Tony
Jebara
s_1
s_2
s_3
s_4
<eos>
hi
,
Tony
Attentive Read
s_4
Embedding for "Tony"
DNN
Selective Read for "Tony"
"Tony"
M
ρ
(c) State Update
h_1
h_2
h_3
h_4
h_5
h_6
h_7
h_8
hello
,
my
name
is
Tony
Jebara
.
(a) Attention-based Encoder-Decoder (RNNSearch)

Decoder: 1. Prediction

- For vocabulary V and unique words in source sequence X
- Instance-specific Vocabulary for source X is $V \cup UNK \cup X$

Given the decoder RNN state $\mathbf{s}_t$ at time t together with $\mathbf{M}$, the probability of generating any target word y_t, is given by the mixture of probabilities as follows:

$$
\begin{aligned}
p(y_t|\mathbf{s}_t, y_{t-1}, \mathbf{c}_t, \mathbf{M}) &= p(y_t, \mathsf{g}|\mathbf{s}_t, y_{t-1}, \mathbf{c}_t, \mathbf{M}) \\
&\quad + p(y_t, \mathsf{c}|\mathbf{s}_t, y_{t-1}, \mathbf{c}_t, \mathbf{M}) \qquad (4)
\end{aligned}
$$

Equations Unrolled:

$$p(y_t|\mathbf{s}_t, y_{t-1}, \mathbf{c}_t, \mathbf{M}) = p(y_t, \mathsf{g}|\mathbf{s}_t, y_{t-1}, \mathbf{c}_t, \mathbf{M}) + p(y_t, \mathsf{c}|\mathbf{s}_t, y_{t-1}, \mathbf{c}_t, \mathbf{M}) \quad (4)$$

g: generative mode

$$p(y_t, \mathsf{g}|\cdot) = \begin{cases} \frac{1}{Z} e^{\psi_g(y_t)}, & y_t \in \mathcal{V} \\ 0, & y_t \in \mathcal{X} \cap \bar{\mathcal{V}} \\ \frac{1}{Z} e^{\psi_g(\mathrm{UNK})} & y_t \notin \mathcal{V} \cup \mathcal{X} \end{cases} \quad (5)$$

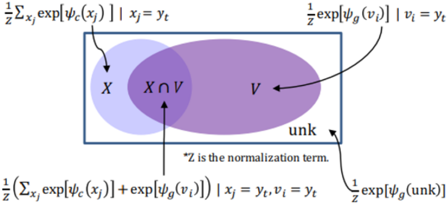
c: copy mode

$$p(y_t, \mathsf{c}|\cdot) = \begin{cases} \frac{1}{Z} \sum_{j:x_j=y_t} e^{\psi_c(x_j)}, & y_t \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

$$\psi_g(y_t = v_i) = \mathbf{v}_i^\top \mathbf{W}_o \mathbf{s}_t, \quad v_i \in \mathcal{V} \cup \mathrm{UNK}$$

$$\psi_c(y_t = x_j) = \sigma\left(\mathbf{h}_j^\top \mathbf{W}_c\right) \mathbf{s}_t, \quad x_j \in \mathcal{X}$$

Normalizing constant $Z = \sum_{v \in \mathcal{V} \cup \{\mathrm{UNK}\}} e^{\psi_g(v)} + \sum_{x \in X} e^{\psi_c(x)}$.


$\frac{1}{Z}\Sigma_{x_j}\exp[\psi_c(x_j)\,] \mid x_j = y_t$
$\frac{1}{Z}\exp[\psi_g(v_i)] \mid v_i = y_t$
X
$X \cap V$
V
unk
$\frac{1}{Z}\left(\Sigma_{x_j}\exp[\psi_c(x_j)] + \exp[\psi_g(v_i)]\right) \mid x_j = y_t, v_i = y_t$
$\frac{1}{Z}\exp[\psi_g(\mathrm{unk})]$

*Z is the normalization term.

Figure 2: The illustration of the decoding probability $p(y_t|\cdot)$ as a 4-class classifier.

Decoder: 2. State Update

- normally $\mathbf{s}_t$ is updated by $\mathbf{s}_{t-1}, y_{t-1}, and\mathbf{c}_t$
- with CopyNET: the y_{t-1} in $y_{t-1} \rightarrow \mathbf{s}_t$ is replaced with:

$$[\mathbf{e}(y_{t-1}); \zeta(y_{t-1})]^\top$$

where $\zeta(y_{t-1}) = \sum_{\tau=1}^{T_S} \rho_{t\tau}\mathbf{h}_\tau$

$$\rho_{t\tau} = \begin{cases} \frac{1}{K} p(x_\tau, \mathsf{c}|\mathbf{s}_{t-1}, \mathbf{M}), & x_\tau = y_{t-1} \\ 0 & \text{otherwise} \end{cases}$$

K = $\sum_{\tau' : x_{\tau'} = y_{t-1}} p(x_{\tau'}, c|\mathbf{s}_{t-1}, \mathbf{M})$

where $\mathbf{e}(y_t 1)$ is the word embedding associated with $y_t 1$, while $\zeta(y_t 1)$ is the weighted sum of hidden states in $\mathbf{M}$ corresponding to y_t

Loss function and Updating

$$\mathcal{L} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\log\left[p(y_t^{(k)}|y_{<t}^{(k)}, X^{(k)})\right]$$

where source sequence $= X^{(N)}$ and target sequence $= Y^{(N)}$

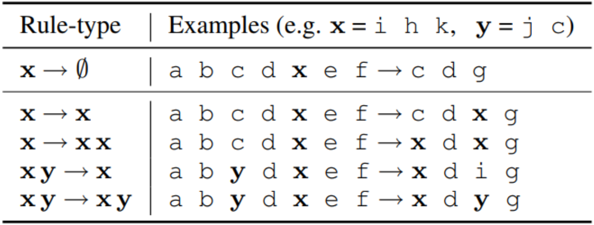
- The network can learn to coordinate the two modes from data
  - if a target word in the source sequence, the copy-mode will contribute to the mixture model, and the gradient will more or less encourage the copy-mode; otherwise, the copy-mode is discouraged due to the competition from the shared normalization term Z

Experiment

1. A synthetic dataset on with simple patterns;
2. A real-world task on text summarization;
3. A dataset for simple single-turn dialogues.

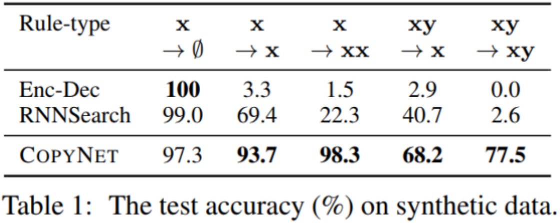
Experiments - Synthetic Dataset

Each rule can further produce a number of instances by replacing the variables with randomly generated subsequences (1 to 15 symbols) from the same vocabulary



Rule-type	Examples (e.g. $\mathbf{x}$ = i h k, $\mathbf{y}$ = j c)
$\mathbf{x} \rightarrow \emptyset$	a b c d **x** e f → c d g
$\mathbf{x} \rightarrow \mathbf{x}$	a b c d **x** e f → c d **x** g
$\mathbf{x} \rightarrow \mathbf{x}\,\mathbf{x}$	a b c d **x** e f → **x** d **x** g
$\mathbf{x}\,\mathbf{y} \rightarrow \mathbf{x}$	a b **y** d **x** e f → **x** d i g
$\mathbf{x}\,\mathbf{y} \rightarrow \mathbf{x}\,\mathbf{y}$	a b **y** d **x** e f → **x** d **y** g

Experiments - Synthetic Dataset



Rule-type	$\mathbf{x}$	$\mathbf{x}$	$\mathbf{x}$	$\mathbf{xy}$	$\mathbf{xy}$
	$\to \emptyset$	$\to \mathbf{x}$	$\to \mathbf{xx}$	$\to \mathbf{x}$	$\to \mathbf{xy}$
Enc-Dec	**100**	3.3	1.5	2.9	0.0
RNNSearch	99.0	69.4	22.3	40.7	2.6
COPYNET	97.3	**93.7**	**98.3**	**68.2**	**77.5**

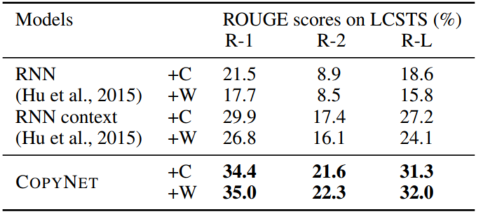
Table 1: The test accuracy (%) on synthetic data.

- Encoder-Decoder (no Attention) → difficulty of representing a long sequence with very high fidelity
- RNNSearch (with Attention) → attention alone seems inadequate for handling the case where strict replication is needed

Experiments - Text Summarization

Automatic text summarization aims to find a condensed representation which can capture the core meaning of the original document

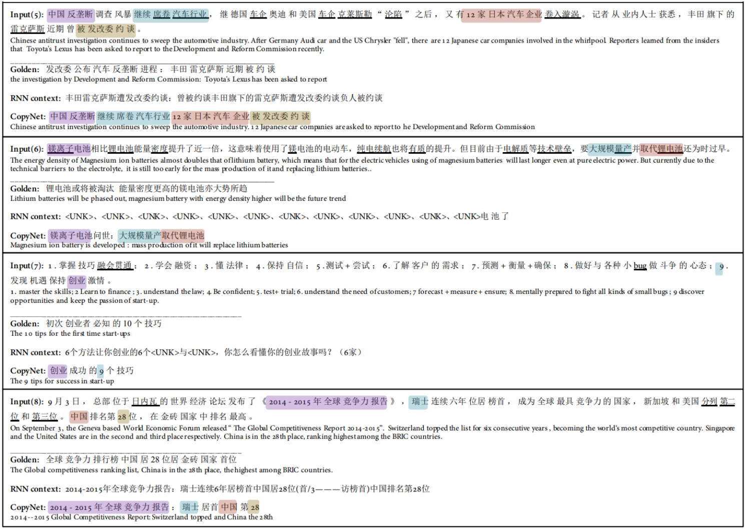
- Dataset: LCSTS dataset (Hu et al., 2015), a large scale dataset for short text summarization in form of (short news, summary).
- model tried on character (+C) and word (+W)
- ROUGE-N: Overlap of N-grams between the system and reference summaries.
- ROUGE - LCN: measures longest matching sequence of words using longest common subsequence.



Models		ROUGE scores on LCSTS (%)		
		R-1	R-2	R-L
RNN	+C	21.5	8.9	18.6
(Hu et al., 2015)	+W	17.7	8.5	15.8
RNN context	+C	29.9	17.4	27.2
(Hu et al., 2015)	+W	26.8	16.1	24.1
COPYNET	+C	**34.4**	**21.6**	**31.3**
	+W	**35.0**	**22.3**	**32.0**

Table 3: Testing performance of LCSTS, where

Experiments - Text Summarization



Input(5): 中国 反垄断 调查 风暴 继续 席卷 汽车行业， 继 德国 车企 奥迪 和 美国 车企 克莱斯勒 “ 沦陷 ” 之后， 又 有 12 家 日本 汽车 企业 卷入漩涡。 记者 从 业内人士 获悉， 丰田 旗下 的 雷克萨斯 近期 曾 被 发改委 约 谈 。
Chinese antitrust investigation continues to sweep the automotive industry. After Germany Audi car and the US Chrysler "fell", there are 12 Japanese car companies involved in the whirlpool. Reporters learned from the insiders that Toyota's Lexus has been asked to report to the Development and Reform Commission recently.

Golden: 发改委 公布 汽车 反垄断 进程： 丰田 雷克萨斯 近期 被 约 谈
the investigation by Development and Reform Commission: Toyota's Lexus has been asked to report

RNN context: 丰田雷克萨斯遭发改委约谈：曾被约谈丰田旗下的雷克萨斯遭发改委约谈负人被约谈

CopyNet: 中国 反垄断 继续 席卷 汽车行业 12 家 日本 汽车 企业 被 发改委 约 谈
Chinese antitrust investigation continues to sweep the automotive industry. 12 Japanese car companies are asked to report to he Development and Reform Commission

Input(6): 镁离子电池相比锂电池能量密度提升了近一倍，这意味着使用了镁电池的电动车，纯电续航也将有质的提升。但目前由于电解质等技术壁垒，要大规模量产并取代锂电池还为时过早。
The energy density of Magnesium ion batteries almost doubles that of lithium battery, which means that for the electric vehicles using of magnesium batteries will last longer even at pure electric power. But currently due to the technical barriers to the electrolyte, it is still too early for the mass production of it and replacing lithium batteries..

Golden: 锂电池或将被淘汰 能量密度更高的镁电池亦大势所趋
Lithium batteries will be phased out, magnesium battery with energy density higher will be the future trend

RNN context: <UNK>、<UNK>、<UNK>、<UNK>、<UNK>、<UNK>、<UNK>、<UNK>、<UNK>、<UNK>、<UNK>、<UNK>电 池 了

CopyNet: 镁离子电池问世：大规模量产取代锂电池
Magnesium ion battery is developed : mass production of it will replace lithium batteries

Input(7): 1 . 掌握 技巧 融会贯通； 2 . 学会 融资； 3 . 懂 法律； 4 . 保持 自信； 5 . 测试 + 尝试； 6. 了解 客户 的 需求； 7. 预测 + 衡量 + 确保； 8 . 做好 与 各种 小 bug 做 斗争 的 心态； 9 . 发现 机遇 保持 创业 激情 。
1. master the skills; 2 Learn to finance ; 3. understand the law; 4. Be confident; 5. test+ trial; 6. understand the need of customers; 7 forecast + measure + ensure; 8. mentally prepared to fight all kinds of small bugs ; 9 discover opportunities and keep the passion of start-up.

Golden: 初次 创业者 必知 的 10 个 技巧
The 10 tips for the first time start-ups

RNN context: 6个方法让你创业的6个<UNK>与<UNK>，你怎么看懂你的创业故事吗？（6家）

CopyNet: 创业 成功 的 9 个 技巧
The 9 tips for success in start-up

Input(8): 9 月 3 日， 总部 位于 日内瓦 的 世界 经济 论坛 发布 了 《 2014 - 2015 年 全球 竞争力 报告 》， 瑞士 连续 六年 位居 榜首， 成为 全球 最具 竞争力 的 国家， 新加坡 和 美国 分列 第二位 和 第三位 。 中国 排名第 28 位， 在 金砖 国家 中 排名 最高 。
On September 3, the Geneva based World Economic Forum released " The Global Competitiveness Report 2014-2015". Switzerland topped the list for six consecutive years , becoming the world's most competitive country. Singapore and the United States are in the second and third place respectively. China is in the 28th place, ranking highest among the BRIC countries.

Golden: 全球 竞争力 排行榜 中国 居 28 位居 金砖 国家 首位
The Global competitiveness ranking list, China is in the 28th place, the highest among BRIC countries.

RNN context: 2014-2015年全球竞争力报告：瑞士连续6年居榜首中国居28位(首/3———访榜首)中国排名第28位

CopyNet: 2014 - 2015 年 全球 竞争力 报告： 瑞士 居首 中国 第 28
2014--2015 Global Competitiveness Report: Switzerland topped and China the 28th

Figure 4: Examples of COPYNET on LCSTS compared with RNN context. Word segmentation is applied on the input, where OOV words are underlined. The highlighted words (with different colors) are those words with copy-mode probability higher than the generate-mode. We also provide literal

Experiments - Text Summarization

1. most words are from copy-mode, but the summary is usually still fluent;
2. COPYNET tends to cover consecutive words in the original document, but it often puts together segments far away from each other, indicating a sophisticated coordination of content-based addressing and location-based addressing;
3. COPYNET handles OOV words really well: it can generate acceptable summary for document with many OOVs, and even the summary itself often contains many OOV words

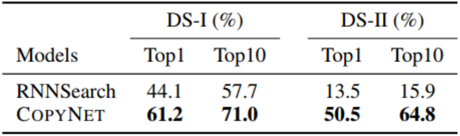
Experiments - Single-turn Dialogue

- Dataset built via a simple dialogue dataset based on the following three instructions
  1. Dialogue instances are collected from Baidu Tieba3 with some coverage of conversations of real life e.g., greeting and sports, etc.
  2. patterns with slots like

     hi, my name is x hi, x

     are mined from the set, with possibly multiple responding patterns to one input.
  3. Similar with the synthetic dataset, we enlarge the dataset by filling the slots with suitable subsequence (e.g. name entities, dates, etc.)
     - Created 2 datasets: DS-I and DS-II
     - the filled substrings for training and testing in DS-II have no overlaps, while in DS-I they are sampled from the same pool

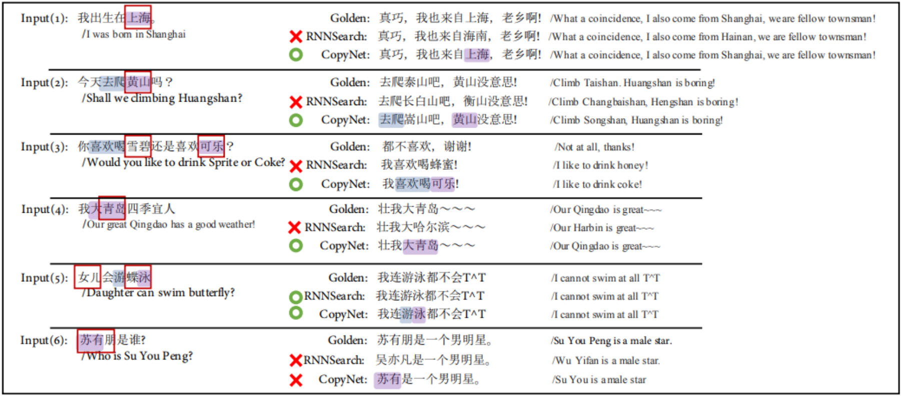
Experiments - Single-turn Dialogue



	DS-I (%)		DS-II (%)	
Models	Top1	Top10	Top1	Top10
RNNSearch	44.1	57.7	13.5	15.9
CopyNet	**61.2**	**71.0**	**50.5**	**64.8**

- Both models estimate respectively the chance of the top-1 or one of top-10 (from beam search) matching the golden.

Experiments - Single-turn Dialogue



Input			Output	Translation
Input(1): 我出生在上海。 /I was born in Shanghai		Golden:	真巧，我也来自上海，老乡啊!	/What a coincidence, I also come from Shanghai, we are fellow townsman!
	✗	RNNSearch:	真巧，我也来自海南，老乡啊!	/What a coincidence, I also come from Hainan, we are fellow townsman!
	○	CopyNet:	真巧，我也来自上海，老乡啊!	/What a coincidence, I also come from Shanghai, we are fellow townsman!
Input(2): 今天去爬黄山吗？ /Shall we climbing Huangshan?		Golden:	去爬泰山吧，黄山没意思!	/Climb Taishan. Huangshan is boring!
	✗	RNNSearch:	去爬长白山吧，衡山没意思!	/Climb Changbaishan, Hengshan is boring!
	○	CopyNet:	去爬嵩山吧，黄山没意思!	/Climb Songshan, Huangshan is boring!
Input(3): 你喜欢喝雪碧还是喜欢可乐？ /Would you like to drink Sprite or Coke?		Golden:	都不喜欢，谢谢!	/Not at all, thanks!
	✗	RNNSearch:	我喜欢喝蜂蜜!	/I like to drink honey!
	○	CopyNet:	我喜欢喝可乐!	/I like to drink coke!
Input(4): 我大青岛四季宜人 /Our great Qingdao has a good weather!		Golden:	壮我大青岛～～～	/Our Qingdao is great~~~
	✗	RNNSearch:	壮我大哈尔滨～～～	/Our Harbin is great~~~
	○	CopyNet:	壮我大青岛～～～	/Our Qingdao is great~~~
Input(5): 女儿会游蝶泳 /Daughter can swim butterfly?		Golden:	我连游泳都不会T^T	/I cannot swim at all T^T
	○	RNNSearch:	我连游泳都不会T^T	/I cannot swim at all T^T
	○	CopyNet:	我连游泳都不会T^T	/I cannot swim at all T^T
Input(6): 苏有朋是谁? /Who is Su You Peng?		Golden:	苏有朋是一个男明星。	/Su You Peng is a male star.
	✗	RNNSearch:	吴亦凡是一个男明星。	/Wu Yifan is a male star.
	✗	CopyNet:	苏有是一个男明星。	/Su You is a male star

Figure 5: Examples from the testing set of DS-II shown as the input text and golden, with the outputs of RNNSearch and CopyNet. Words in red rectangles are unseen in the training set. The highlighted words (with different colors) are those words with copy-mode probability higher than the generate-mode. Green cirles (meaning correct) and red cross (meaning incorrect) are given based on human judgment on whether the response is appropriate.