The 2021 Conference on Empirical Methods in Natural Language Processing

SOM-NCSCM: An Efficient Neural Chinese Sentence Compression Model Enhanced with Self-Organizing Map

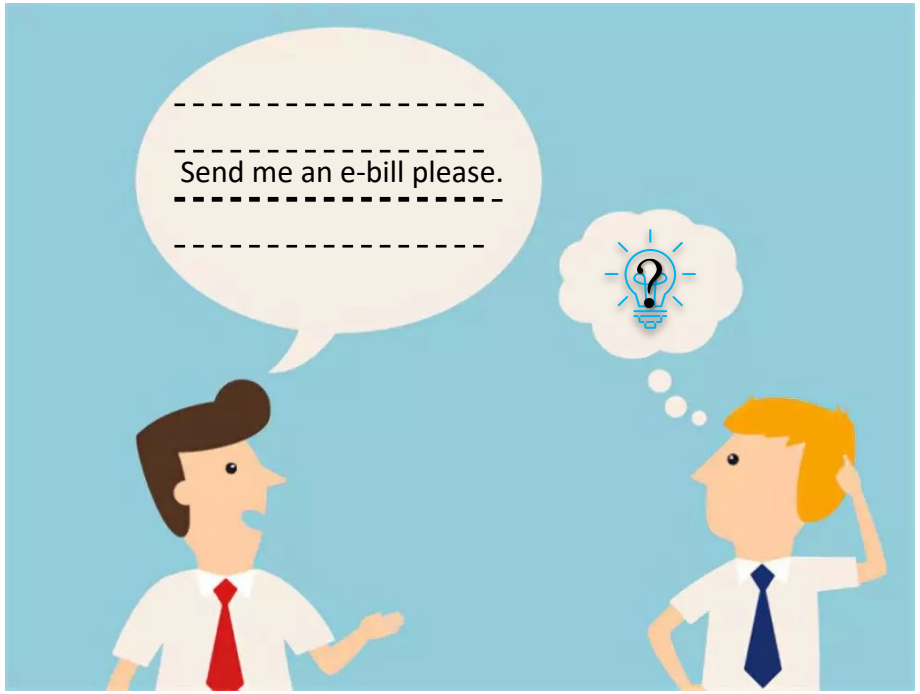
Kangli Zi[1,2], Shi Wang[1], Yanan Cao[3], Yu Liu[1,2], Jicun Li[1,2], Cungen Cao[1]

[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Institute of Information Engineering, Chinese Academy of Sciences

Outline

- Motivation
- Dataset
- Models
- Results

What is Sentence Compression (SC)?



- The **SC** aims to **shorten verbose sentences into concise ones**.
- Most approaches have treated it as a **deletion-based processing** and formalized it as a **sequence labeling problem**.
- The SC can benefit several real applications, such as automatic title generation (Zhang et al., 2012; Wang et al., 2018), information extraction, opinion mining (Feng et al., 2010), machine translation (Liet al., 2020), and question answering systems.

- Motivation
- Dataset
- Models
- Results

Problems in Chinese Sentence Compression

1. Lacking corpora of Chinese parallel data which are necessary for training and evaluating Chinese SC models.
   - Translating the English SC corpora into Chinese, or collecting data from news websites.
   - But those data are mostly written in **Chinese formal expressions** and **not made publicly available** or exactly extraction-based.
2. High cost of annotating Chinese data.
3. Current Chinese SC approaches are more unsupervised-method-based.
   - Using heuristic rules ⟶ hard to be reused
   - Using statistical probabilities, like the TF-IDF ⟶ lack flexibility and capacity

Contributions

1. Lacking corpora of Chinese parallel data which are necessary for training and evaluating Chinese SC models.

1. A Chinese SC dataset
   - Chinese colloquial sentences from a real-life QA system in the telecommunication domain.

2. High cost of annotating Chinese data.

3. Current Chinese SC approaches are more unsupervised-method-based.
   - Using heuristic rules ⟶ hard to be reused
   - Using statistical probabilities, like the TF-IDF ⟶ lack flexibility and capacity

Contributions

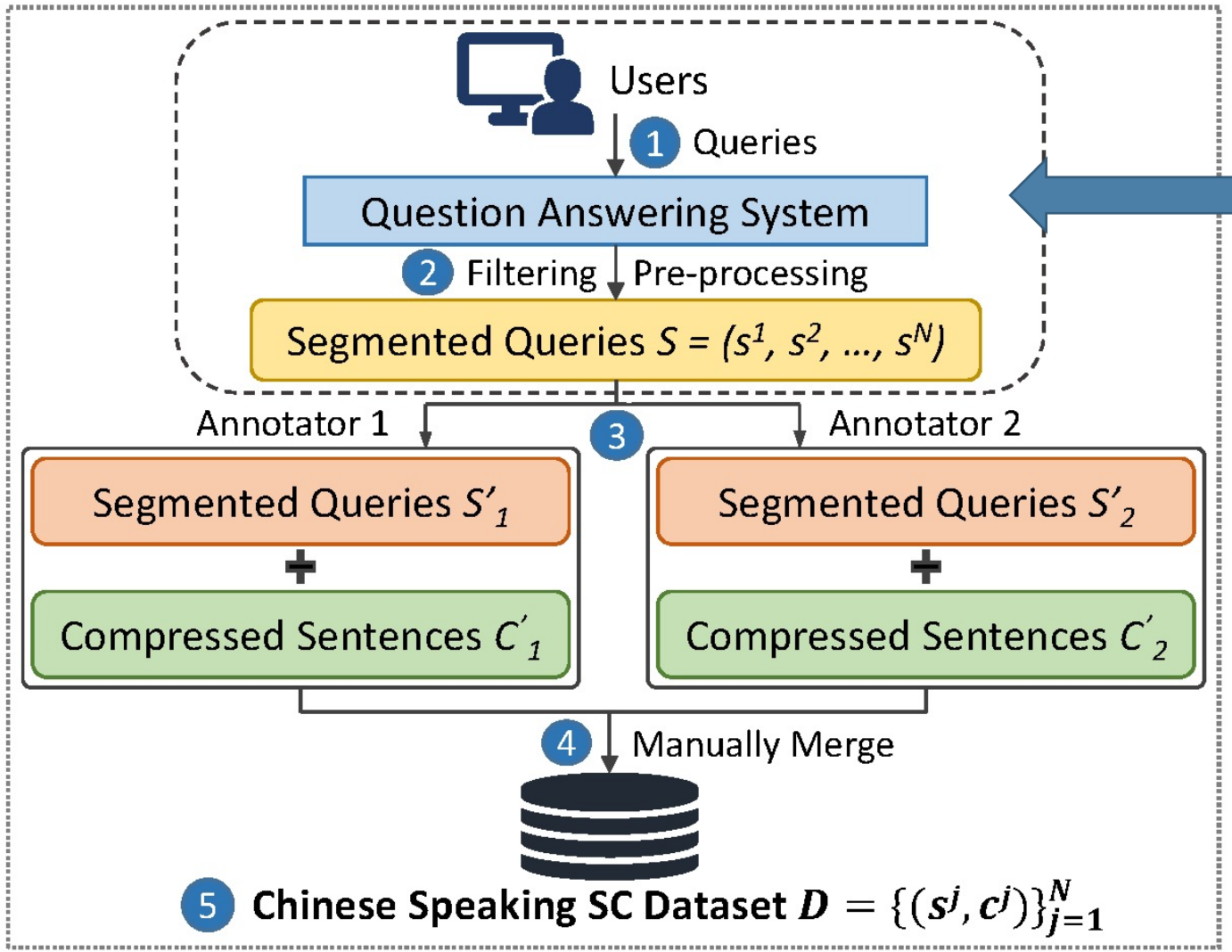
1. Lacking corpora of Chinese parallel data which are necessary for training and evaluating Chinese SC models.

1. A Chinese SC dataset
   - Chinese colloquial sentences from a real-life QA system in the telecommunication domain.

2. An entire neural Chinese SC model — SOM-NCSCM
   - Using a neural cluster method to enhance the performance of the neural Chinese SC model.

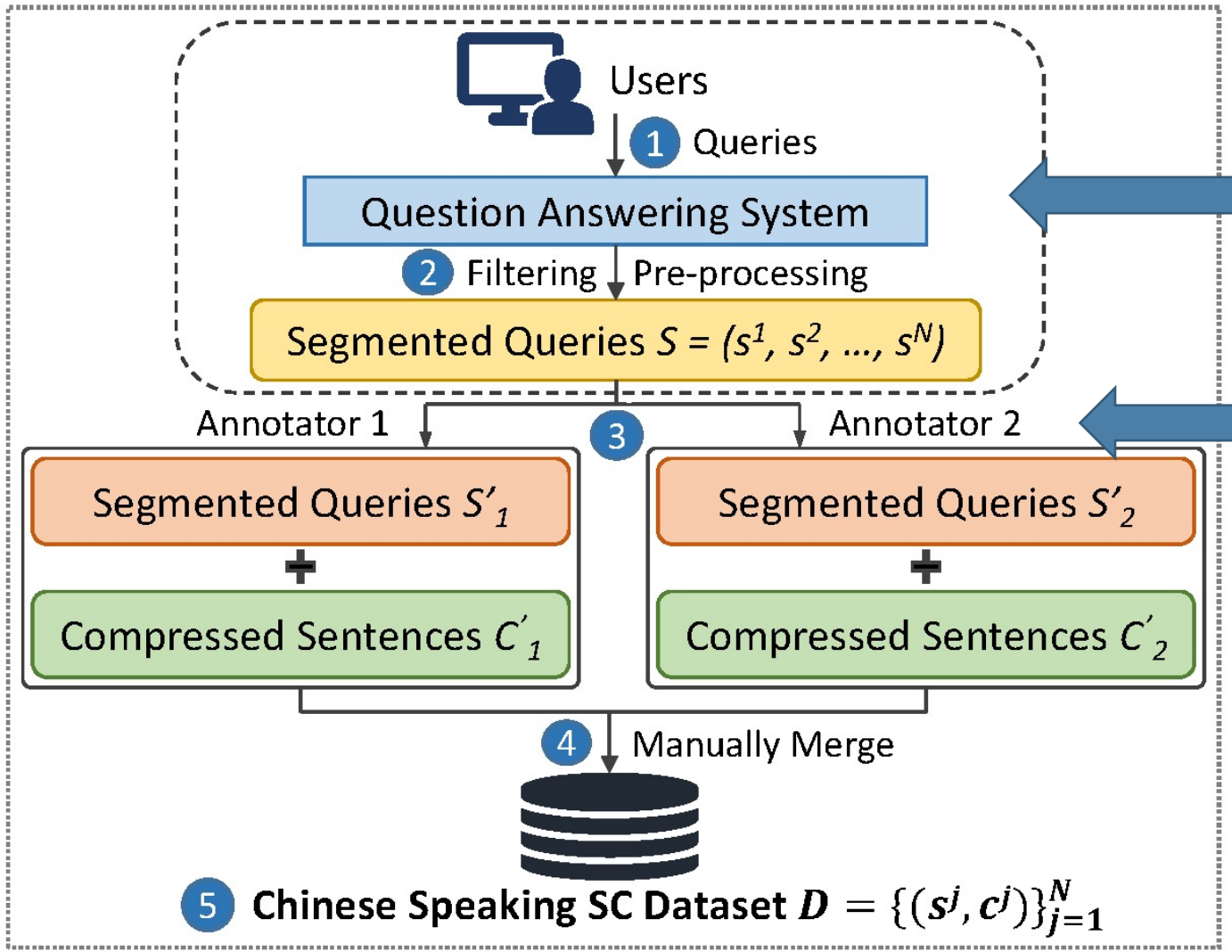
- Motivation
- Dataset
- Models
- Results

Data Collection and Annotation



✓ Data Source

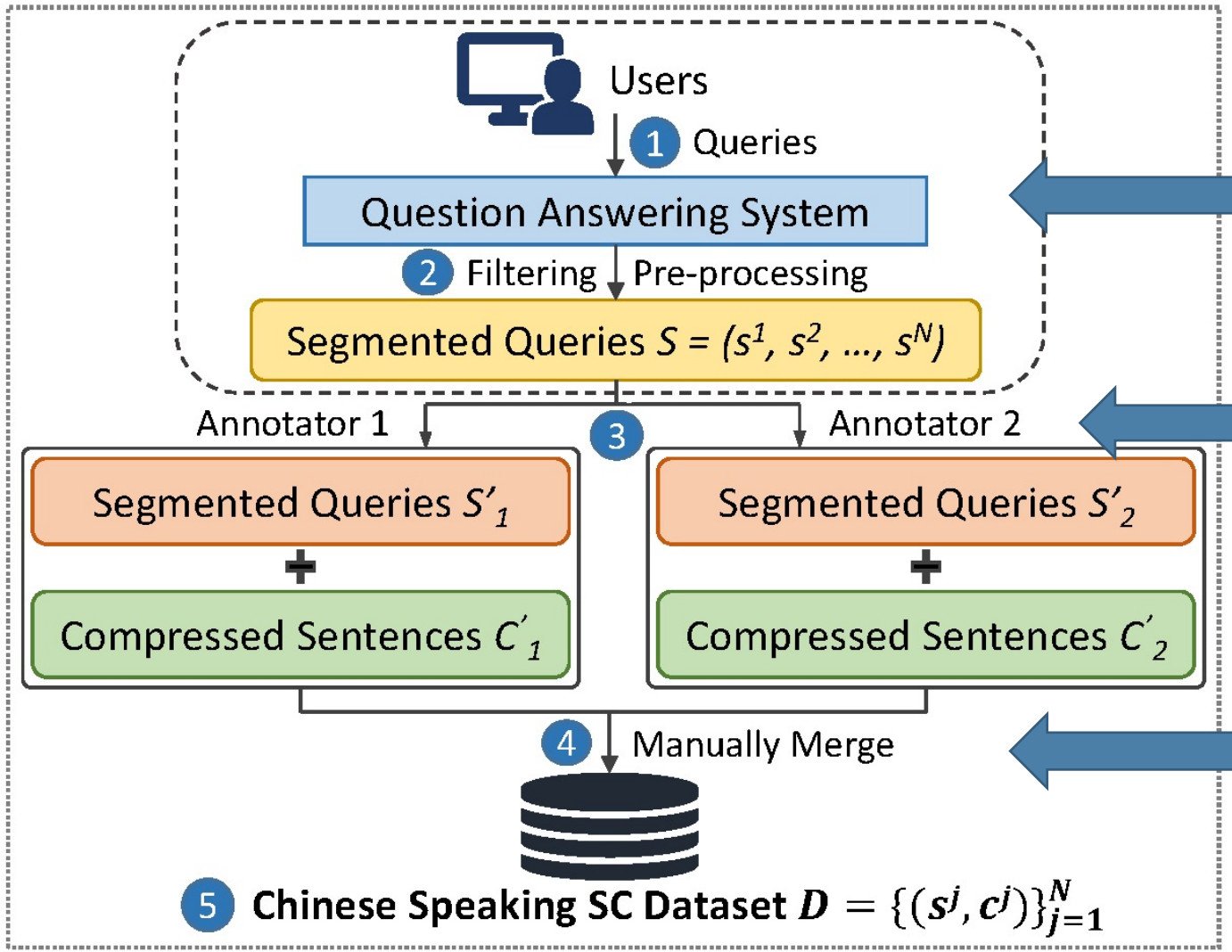
Data Collection and Annotation



✓ Data Source

✓ Data Annotation

Data Collection and Annotation



✓ Data Source

✓ Data Annotation

✓ Merge Data

Data Evaluation and Description

✓ Data Evaluation

- Automatic and quantitative evaluation metrics
  - · A list of high-frequency words in the whole data, except the common stop words in Chinese.
  - · A list of words left in real-time compressed sentences together with their labels and the corresponding frequencies.
- The Cohen's Kappa coefficient ⟶ **The inter-annotator agreement**
  - · The Cohen's unweighted k = 0.623 ⟶ A substantial level
- The Hopkins statistic ⟶ The **cluster tendency** of our data
  - · The value is around 0.719 – 0.726 ⟶ A high tendency to be clustered

Data Evaluation and Description

- ✓ Data Evaluation
- ✓ Data Description

<table>
<tr><td colspan="2">Total Number</td><td colspan="3">3300 (original query, compressed sentence) pairs</td></tr>
<tr><td colspan="2">Compression Ratio (CR)</td><td colspan="3">0.709</td></tr>
<tr><td rowspan="2">Total Words</td><td>Origin: 2,552</td><td rowspan="2">Sentence Length</td><td rowspan="2">Origin</td><td>2 - 48 (characters)</td></tr>
<tr><td rowspan="2">Compression: 2,291</td><td>1 - 28 (words)</td></tr>
<tr><td rowspan="2">Compression</td><td>2 - 37 (characters)</td></tr>
<tr><td>1 - 18 (words)</td></tr>
</table>

Note: $CR = \frac{The\ number\ of\ left\ words\ in\ the\ compressed\ sentences}{The\ total\ number\ of\ word\ in\ the\ original\ sentences}$

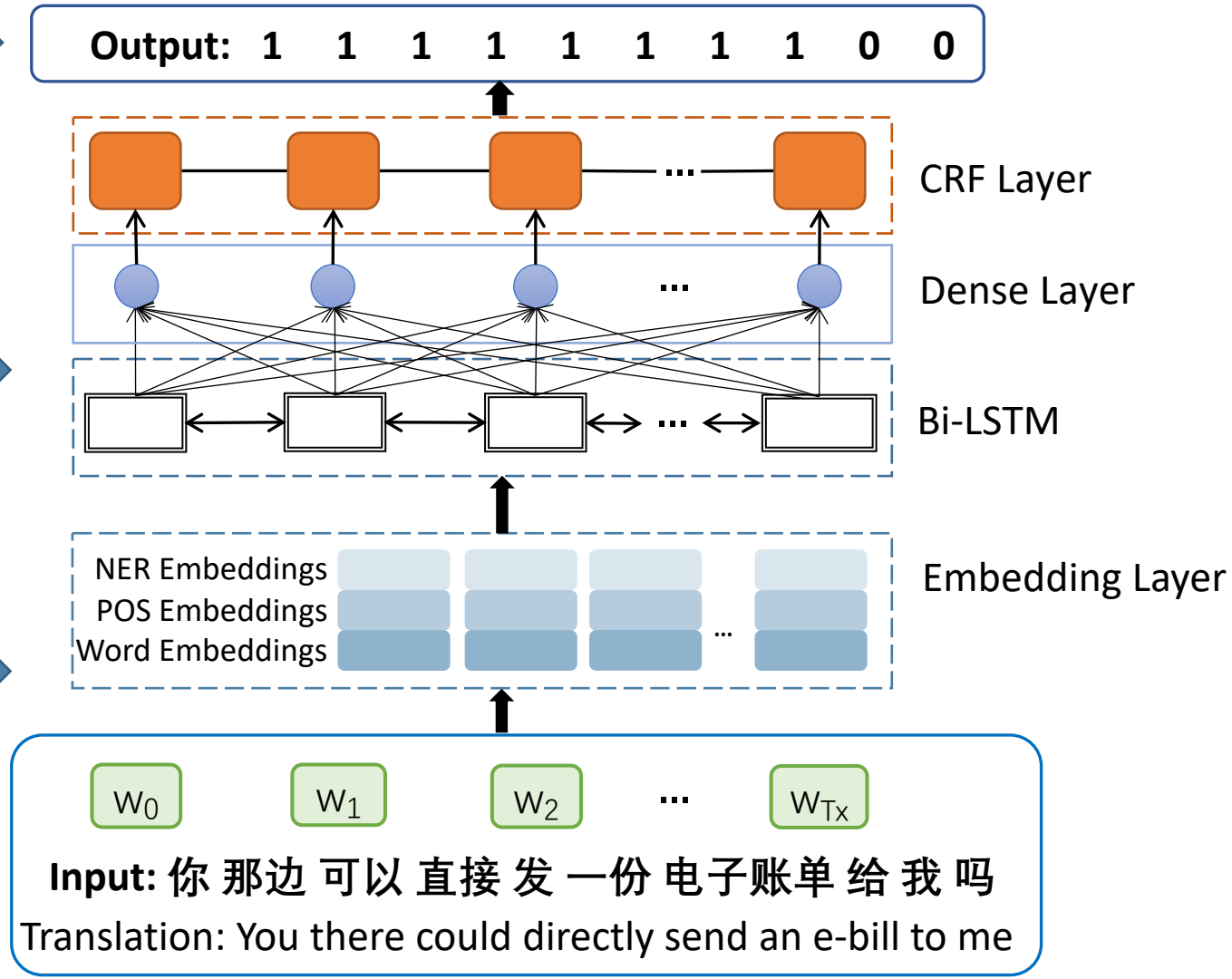
- Motivation
- Dataset
- Models
- Results

Baseline Model

Objective

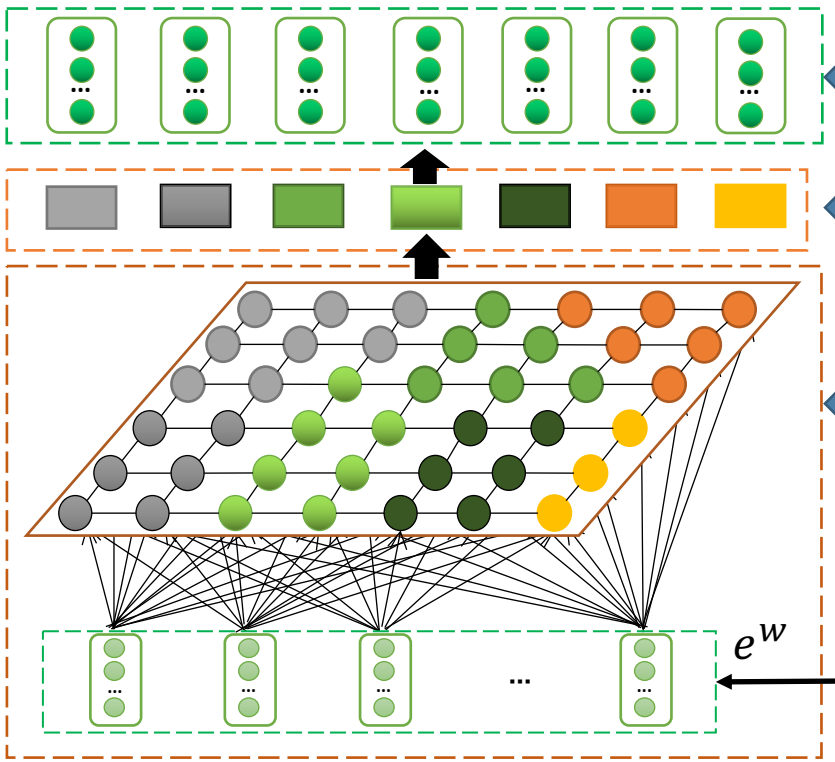
A commonly-used neural sequence labeling model: Bi-LSTM+CRF

Features: named entities (NE) features, part-of-speech (POS) features.



Pre-trained SOM

- For **better exploring the implication of our data and relieving the degree of data sparsity**, we apply a neural cluster model (the Self-Organizing Map, SOM) to obtain the cluster information of our data.



Cluster index embeddings

$z_s = som(e^w; \theta_s)$ → a mono-dimensional index

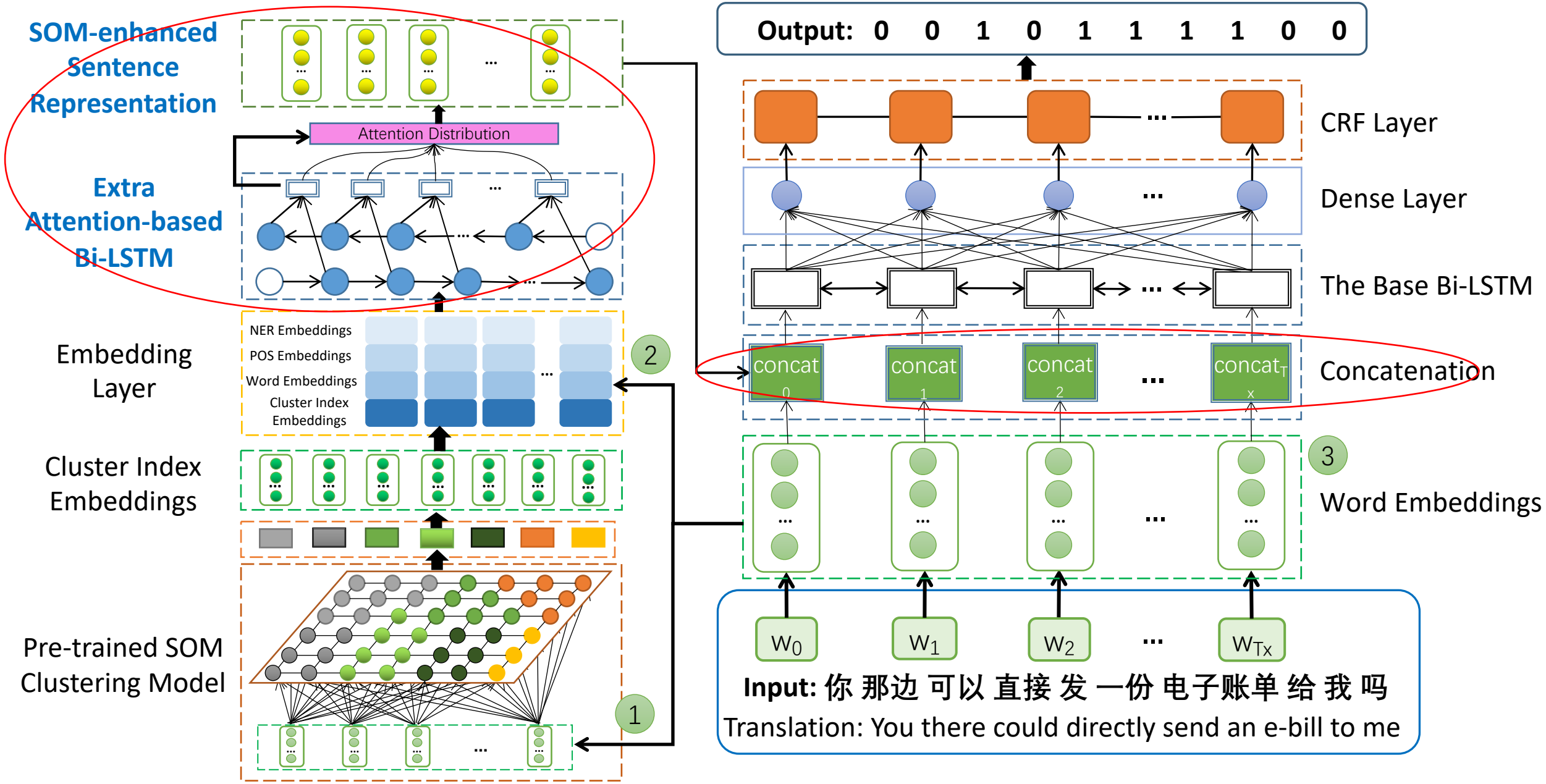
SOM: a feed-forward structure with a square output layer

w_0 w_1 w_2 ... w_{Tx}

Input: 你 那边 可以 直接 发 一份 电子账单 给 我 吗

Translation: You there could directly send an e-bill to me

SOM-NCSCM


SOM-enhanced Sentence Representation
Extra Attention-based Bi-LSTM
Attention Distribution
Embedding Layer
NER Embeddings
POS Embeddings
Word Embeddings
Cluster Index Embeddings
Cluster Index Embeddings
Pre-trained SOM Clustering Model
2
1
3
Output: 0 0 1 0 1 1 1 1 0 0
CRF Layer
Dense Layer
The Base Bi-LSTM
concat 0
concat 1
concat 2
concat T x
Concatenation
Word Embeddings
W0
W1
W2
WTx
Input: 你 那边 可以 直接 发 一份 电子账单 给 我 吗
Translation: You there could directly send an e-bill to me

- Motivation
- Dataset
- Models
- Results

Main Results

Models	F1	BLEU1	BLEU2	BLEU3	BLEU4	CR
Chinese BERT-based model	85.785	73.442	64.253	58.295	53.911	0.833
Baseline	88.466	82.996	76.179	71.210	67.449	0.690
w/ SOM direct	88.228	81.162	76.214	70.804	66.645	0.649
w/o NE&POS	87.872	81.250	75.155	68.773	64.124	0.712
w/o NE&POS w/ SOM direct	87.991	82.313	75.948	69.861	64.431	0.702
K-Means + NCSCM	89.061	**84.238**	77.995	72.299	67.333	0.672
GMM + NCSCM	88.704	82.893	**79.201**	73.763	68.489	0.673
SOM-NCSCM	**89.655**	84.000	78.221	**74.766**	**70.116**	0.683

Training set: 3,000
Development set: 150
Test set: 150

Main Results

Models	F1	BLEU1	BLEU2	BLEU3	BLEU4	CR
Chinese BERT-based model	85.785	73.442	64.253	58.295	53.911	0.833
Baseline	88.466	82.996	76.179	71.210	67.449	0.690
w/ SOM direct	88.228	81.162	76.214	70.804	66.645	0.649
w/o NE&POS	87.872	81.250	75.155	68.773	64.124	0.712
w/o NE&POS w/ SOM direct	87.991	82.313	75.948	69.861	64.431	0.702
K-Means + NCSCM	89.061	**84.238**	77.995	72.299	67.333	0.672
GMM + NCSCM	88.704	82.893	**79.201**	73.763	68.489	0.673
SOM-NCSCM	**89.655**	84.000	78.221	**74.766**	**70.116**	0.683

1. The scores of the fine-tuned Chinese BERT-based model don't reach fascinating points.
   · CR = 0.833 ⟶ Tends to **retain more words**.
   · It's easy to be **over-fitting**, due to the complex structure.

Main Results

Models	F1	BLEU1	BLEU2	BLEU3	BLEU4	CR
Chinese BERT-based model	85.785	73.442	64.253	58.295	53.911	0.833
Baseline	88.466	82.996	76.179	71.210	67.449	0.690
w/ SOM direct	88.228	81.162	76.214	70.804	66.645	0.649
w/o NE&POS	87.872	81.250	75.155	68.773	64.124	0.712
w/o NE&POS w/ SOM direct	87.991	82.313	75.948	69.861	64.431	0.702
K-Means + NCSCM	89.061	**84.238**	77.995	72.299	67.333	0.672
GMM + NCSCM	88.704	82.893	**79.201**	73.763	68.489	0.673
SOM-NCSCM	**89.655**	84.000	78.221	**74.766**	**70.116**	0.683

2. All the models employed with lexical features (NE and POS features) perform better than those without them.
   · **The effectiveness of the lexical features**.

Main Results

Models	F1	BLEU1	BLEU2	BLEU3	BLEU4	CR
Chinese BERT-based model	85.785	73.442	64.253	58.295	53.911	0.833
Baseline	88.466	82.996	76.179	71.210	67.449	0.690
w/ SOM direct	88.228	81.162	76.214	70.804	66.645	0.649
w/o NE&POS	87.872	81.250	75.155	68.773	64.124	0.712
w/o NE&POS w/ SOM direct	87.991	82.313	75.948	69.861	64.431	0.702
K-Means + NCSCM	89.061	**84.238**	77.995	72.299	67.333	0.672
GMM + NCSCM	88.704	82.893	**79.201**	73.763	68.489	0.673
SOM-NCSCM	**89.655**	84.000	78.221	**74.766**	**70.116**	0.683

3. Adding the cluster index features generally helps.

Main Results

Models	F1	BLEU1	BLEU2	BLEU3	BLEU4	CR
Chinese BERT-based model	85.785	73.442	64.253	58.295	53.911	0.833
Baseline	88.466	82.996	76.179	71.210	67.449	0.690
w/ SOM direct	88.228	81.162	76.214	70.804	66.645	0.649
w/o NE&POS	87.872	81.250	75.155	68.773	64.124	0.712
w/o NE&POS w/ SOM direct	87.991	82.313	75.948	69.861	64.431	0.702
K-Means + NCSCM	89.061	**84.238**	77.995	72.299	67.333	0.672
GMM + NCSCM	88.704	82.893	**79.201**	73.763	68.489	0.673
SOM-NCSCM	**89.655**	84.000	78.221	**74.766**	**70.116**	0.683

3. Adding the cluster index features causes the performance to drop a little, compared to the standard baseline model.
   · Noises and sparsity problem.
   · The amount of dataset is not so large to make **a trade-off between fixing clustering mistakes and learning from cluster index features.**

Main Results

Models	F1	BLEU1	BLEU2	BLEU3	BLEU4	CR
Chinese BERT-based model	85.785	73.442	64.253	58.295	53.911	0.833
Baseline	88.466	82.996	76.179	71.210	67.449	0.690
w/ SOM direct	88.228	81.162	76.214	70.804	66.645	0.649
w/o NE&POS	87.872	81.250	75.155	68.773	64.124	0.712
w/o NE&POS w/ SOM direct	87.991	82.313	75.948	69.861	64.431	0.702
K-Means + NCSCM	89.061	**84.238**	77.995	72.299	67.333	0.672
GMM + NCSCM	88.704	82.893	**79.201**	73.763	68.489	0.673
SOM-NCSCM	**89.655**	84.000	78.221	**74.766**	**70.116**	0.683

4. The efficient and feasible structure of our neural Chinese SC model.

→ The advantageous ability of the SOM, and the whole SOM-NCSCM can alleviate the effect of the shortage of parallel data while also make better use of similarity among data.

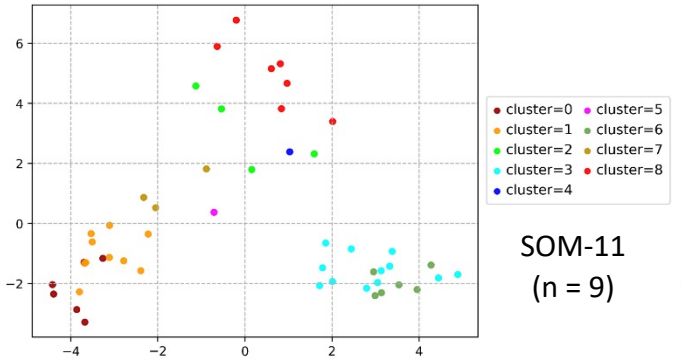
Analyses of Cluster Results

✓ Experiments on SOM parameters

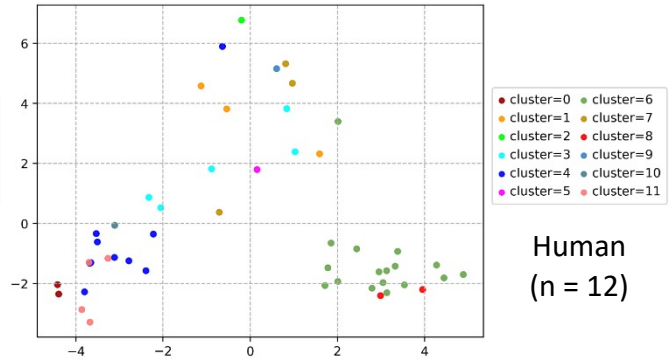
Our SOM-NCSCM					
SOM size	9 × 9	10 × 10	11 × 11	12 × 12	13 × 13
F1	89.299	88.704	**89.655**	88.466	88.823
BLEU1	83.908	81.227	**84.000**	83.206	81.132
BLEU2	**79.156**	76.425	78.221	77.663	75.552
BLEU3	74.263	70.606	**74.766**	73.876	71.400
BLEU4	**70.547**	66.207	70.116	70.016	66.470
CR	0.688	0.649	0.683	0.680	0.710
Number of Clusters	81	100	121	144	169

➢ **The Silhouette Coefficient scores and Calinski-Harabasz scores of all experimental models and human judgement on 50 randomly selected query sentences.**

✓ Cluster results of 50 Randomly-selected query sentences



SOM-11
(n = 9)



Human
(n = 12)

Model	Silhouette Coefficient Score	Calinski-Harabasz Score
K-Means	0.032	2.451
GMM	0.007	2.205
SOM-9	0.026	2.597
SOM-10	0.045	2.788
SOM-11	**0.071**	**4.066**
SOM-12	0.053	**3.002**
SOM-13	0.035	2.838
Human	**0.065**	2,945

➢ Manual judgement criteria: the user demands, the sentence length, the structure of original query sentences.

Case Studies

Original Query Sntence	你/那边/可以/直接/发/一份/电子账单/给/我/吗	你/要/给/记录/[编号]/号/客服代表/挂/我电话/的/问题
Translation in English	You there could directly send an e-bill to me	You should record for (me) the matter that the No. [number] customer service representative hangs up on me
Gold Compressed Sentence	发/一份/电子账单/给/我 Send an e-bill to me	[编号]/号/客服代表/挂/我电话 The No. [number] customer service representative hangs up on me
Chinese BERT-based model	{边}/可以/直接/发/{份}/电子/账单/给/我 {There} could directly send {an} e-bill to me	记录/[编号]/号/客服/代表/挂/我/电话/问题 Record the matter that the No. [number] customer service representative hangs up on me
Baseline	你/那边/可以/直接/发/一份/电子账单/给 You there could directly send an e-bill to	[编号]/号/挂/我电话/问题 The matter that the No. [number] hangs up on me
K-Means + NCSCM	发/一份/电子账单/给 Send an e-bill to	挂/我电话/问题 The matter that hangs up on me
GMM + NCSCM	可以/发/一份/电子账单 Could send an e-bill	号/挂/我电话/问题 The matter that the No. hangs up on me
SOM-NCSCM	可以/发/一份/电子账单/给 Could send an e-bill to	[编号]/号/挂/我电话/问题 The matter that the No. [number] hangs up on me

- There's some degree of gap between **getting exact matched results** with strict gold compressed sentences and producing acceptable outputs.

Case Studies

Original Query Sntence	你/那边/可以/直接/发/一份/电子账单/给/我/吗	你/要/给/记录/[编号]/号/客服代表/挂/我电话/的/问题
Translation in English	You there could directly send an e-bill to me	You should record for (me) the matter that the No. [number] customer service representative hangs up on me
Gold Compressed Sentence	发/一份/电子账单/给/我 Send an e-bill to me	[编号]/号/客服代表/挂/我电话 The No. [number] customer service representative hangs up on me
Chinese BERT-based model	{边}/可以/直接/发/{份}/电子/账单/给/我 {There} could directly send {an} e-bill to me	记录/[编号]/号/客服/代表/挂/我/电话/问题 Record the matter that the No. [number] customer service representative hangs up on me
Baseline	你/那边/可以/直接/发/一份/电子账单/给 You there could directly send an e-bill to	[编号]/号/挂/我电话/问题 The matter that the No. [number] hangs up on me
K-Means + NCSCM	发/一份/电子账单/给 Send an e-bill to	挂/我电话/问题 The matter that hangs up on me
GMM + NCSCM	可以/发/一份/电子账单 Could send an e-bill	号/挂/我电话/问题 The matter that the No. hangs up on me
SOM-NCSCM	可以/发/一份/电子账单/给 Could send an e-bill to	[编号]/号/挂/我电话/问题 The matter that the No. [number] hangs up on me

- The fine-tuned Chinese BERT-based model keeps **the most characters** in its outputs.
  There are even some **incorrectly labeled words**.
  - Such as “那边” (there) and “一份” (an).

Case Studies

Original Query Sntence	你/那边/可以/直接/发/一份/电子账单/给/我/吗	你/要/给/记录/[编号]/号/客服代表/挂/我电话/的/问题
Translation in English	You there could directly send an e-bill to me	You should record for (me) the matter that the No. [number] customer service representative hangs up on me
Gold Compressed Sentence	发/一份/电子账单/给/我 Send an e-bill to me	[编号]/号/客服代表/挂/我电话 The No. [number] customer service representative hangs up on me
Chinese BERT-based model	{边}/可以/直接/发/{份}/电子/账单/给/我 {There} could directly send {an} e-bill to me	记录/[编号]/号/客服/代表/挂/我/电话/问题 Record the matter that the No. [number] customer service representative hangs up on me
Baseline	你/那边/可以/直接/发/一份/电子账单/给 You there could directly send an e-bill to	[编号]/号/挂/我电话/问题 The matter that the No. [number] hangs up on me
K-Means + NCSCM	发/一份/电子账单/给 Send an e-bill to	挂/我电话/问题 The matter that hangs up on me
GMM + NCSCM	可以/发/一份/电子账单 Could send an e-bill	号/挂/我电话/问题 The matter that the No. hangs up on me
SOM-NCSCM	可以/发/一份/电子账单/给 Could send an e-bill to	[编号]/号/挂/我电话/问题 The matter that the No. [number] hangs up on me

- Mistaken deletion of the keyword “客服代表 (customer service representative)” in most of models.
  - · A **domain-specific** proper noun
  - · An unknown word in the pre-trained word embeddings
  - · It’s unrecognizable as a whole word phrase for the tools to label NE and POS tags.
- The Chinese BERT-based model can hold the correct word tokens but in a **different segmentation granularity**.

Conclusion

1. We construct **a Chinese SC dataset**, composed of Chinese colloquial sentences, from a real-life QA system in the telecommunication domain.
2. We build **several neural baseline models** and propose our **SOM-NCSCM** for Chinese SC task.

- We'll work on constructing a **larger Chinese SC dataset** and take more measures to ensure the quality of the dataset.
- We also plan to explore **domain-specific knowledge** and **other neural-network techniques** to deal with the Chinese SC task in the future.

Dataset is available at *https://github.com/Zikangli/SOM-NCSCM*.

The 2021 Conference on Empirical Methods in Natural Language Processing

SOM-NCSCM: An Efficient Neural Chinese Sentence Compression Model Enhanced with Self-Organizing Map

Thanks for your listening!

Kangli Zi, Shi Wang, Yanan Cao, Yu Liu, Jicun Li, Cungen Cao

中国科学院大学 University of Chinese Academy of Sciences

ICT 中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

中国科学院信息工程研究所 INSTITUTE OF INFORMATION ENGINEERING,CAS