Multi-task Learning for Sentence Compression

English Reading Group

Yang Zhao

Improving sentence compression by learning to predict gaze

Sigrid Klerke, Yoav Goldberg Bar-Ilan, and Anders Søgaard

NAACL, 2016, best paper

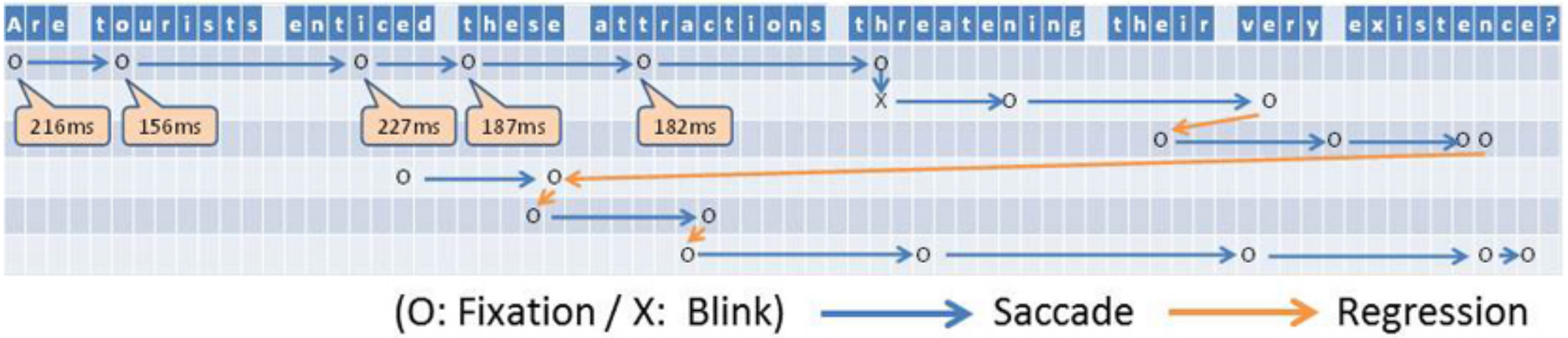
http://www.aclweb.org/anthology/N16-1179

Motivation of using eye-tracking recordings for improving sentence compression

Motivation-1: Sentence compression is the task of automatically making sentences easier to process by shortening them.

Motivation-2: Eye-tracking measures such as first-pass reading time and time spent on regressions, i.e., during second and later passes over the text, are known to correlate with perceived text difficulty (Rayner et al., 2012).

First-pass reading time and regressions duration



Gaze during reading

Words	First Pass	Regressions
Are	4	4
tourists	2	0
enticed	3	0
by	4	0
these	2	0
attractions	3	0
threatening	3	3
their	5	0
very	3	3
existence	3	5
?	3	5

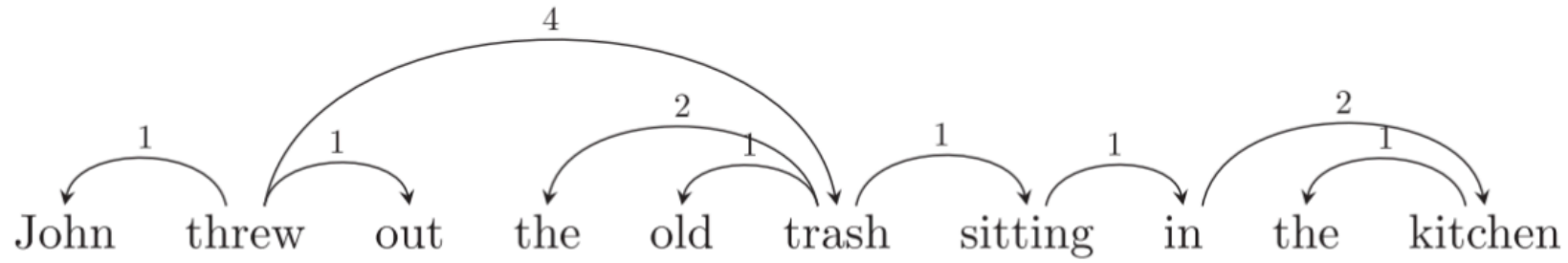
Figure 1: Example sentence from the Dundee Corpus

0: measure = 0 or
1: measure < 1 SD below reader's average or
2: measure < .5 SD below reader's average or
3: measure < .5 above reader's average or
4: measure > .5 SD above reader's average or
5: measure > 1 SD above reader's average

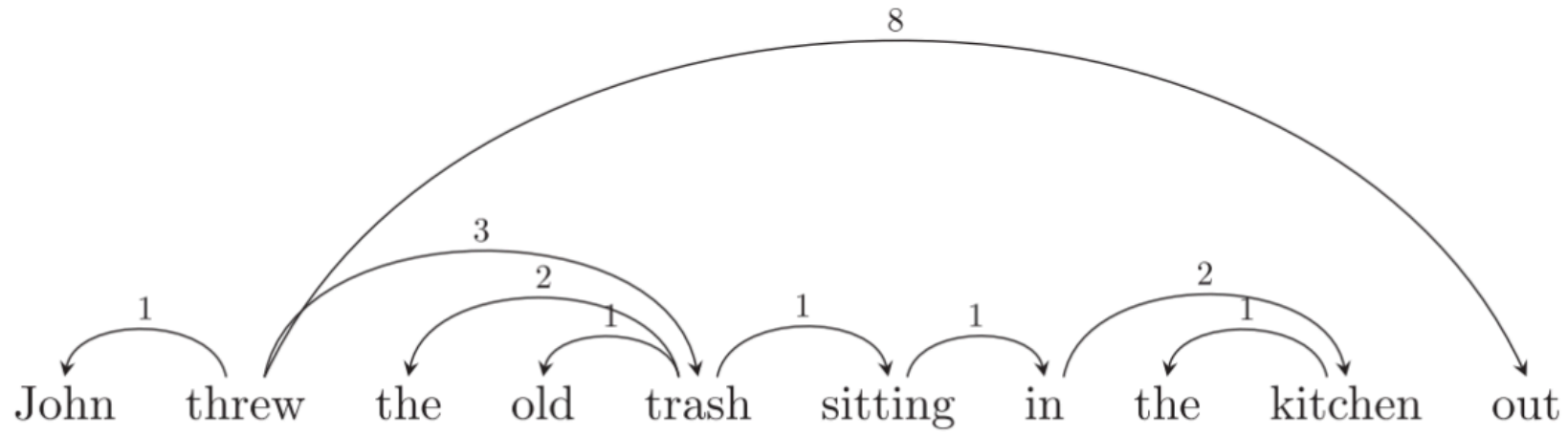
- Are tourists enticed by these attractions threatening their very existence ?
  - First-pass reading time: the total time spent reading a word first time it is fixated. And it correlates with the following factors:
    - Word length
    - Frequency
    - Ambiguity
  - Regression duration: the total time spent fixating a word after the gaze has already left it once.
    - semantic confusion and contradiction
    - incongruence
    - **syntactic complexity**

Syntactic Complexity - Dependency Locality

The cost of integrating two elements depends on the distance between them



Sentence C: Total dependency length = 14



Sentence D: Total dependency length = 20

Gaze during reading

Words	First Pass	Regressions
Are	4	4
tourists	2	0
enticed	3	0
by	4	0
these	2	0
attractions	3	0
threatening	3	3
their	5	0
very	3	3
existence	3	5
?	3	5

Figure 1: Example sentence from the Dundee Corpus

0: measure = 0 or
1: measure < 1 SD below reader's average or
2: measure < .5 SD below reader's average or
3: measure < .5 above reader's average or
4: measure > .5 SD above reader's average or
5: measure > 1 SD above reader's average

- Are tourists enticed by these attractions threatening their very existence ?
  - First-pass reading time: the total time spent reading a word first time it is fixated. And it correlates with the following factors:
    - Word length
    - Frequency
    - Ambiguity
  - Regression duration: the total time spent fixating a word after the gaze has already left it once.
    - semantic confusion and contradiction
    - incongruence
    - **syntactic complexity**

Proposed two models: (1) MULTI-TASK-LSTM and (2) CASCADED-LSTM

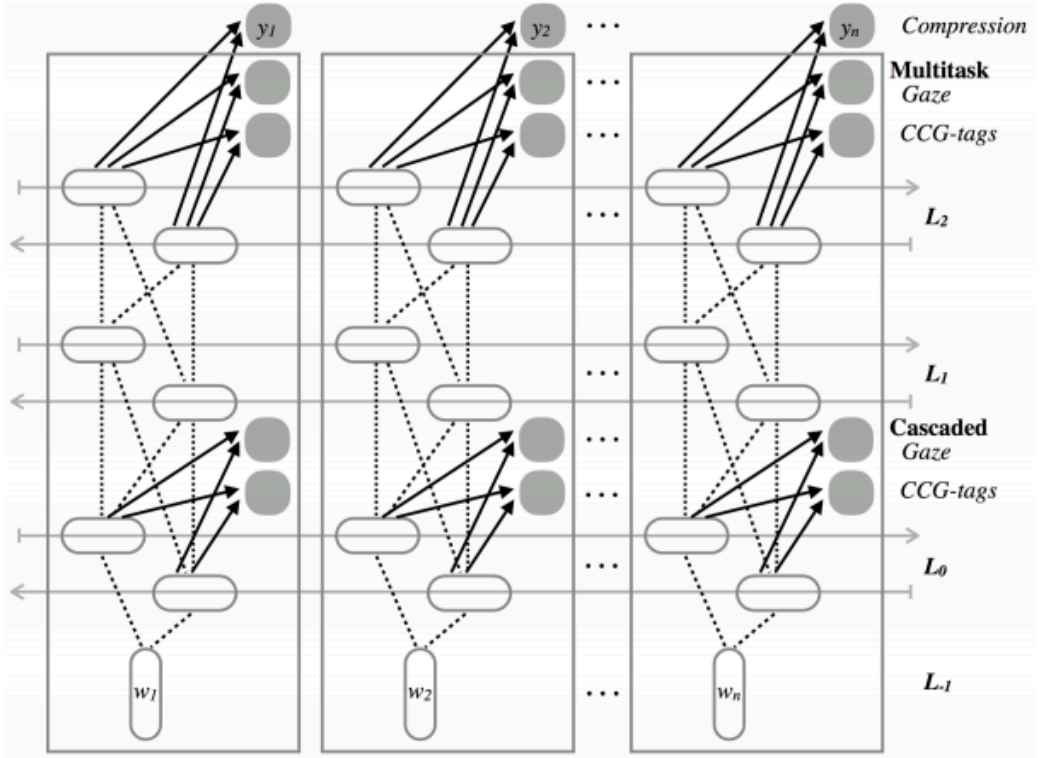


Figure 2: Multitask and cascaded bi-LSTMs for sentence compression. Layer L_{-1} contain pre-trained embeddings. Gaze prediction and CCG-tag prediction are auxiliary training tasks, and loss on all tasks are propagated back to layer L_0.

MULTI-TASK-LSTM

CASCADED-LSTM

Difference: where to add the gaze information

Input: word

Output:

- Compression label: {0,1}
- Gaze label: {0,1,2,3,4,5}
- CCG-tag: comes from CCGbank

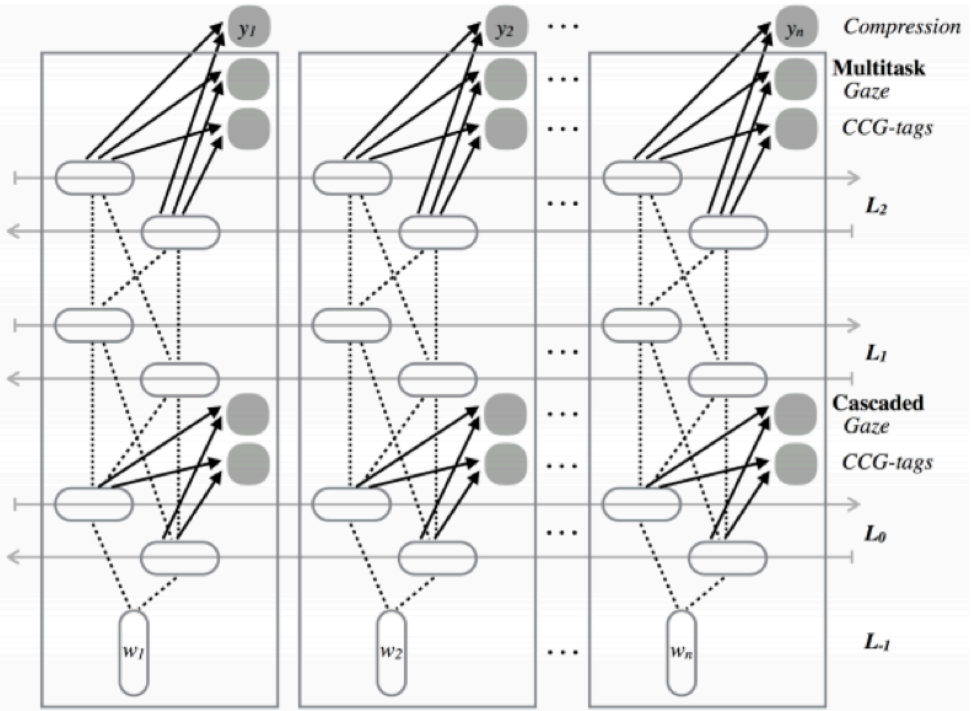
Data

	Sents	Sent.len	Type/token	Del.rate
		TRAINING		
ZIFF-DAVIS	1000	20	0.22	0.59
BROADCAST	880	20	0.21	0.27
GOOGLE	8000	24	0.17	0.87
		TEST		
ZIFF-DAVIS	32	21	0.55	0.47
BROADCAST	412	19	0.27	0.29
GOOGLE	1000	25	0.42	0.87

Table 2: Dataset characteristics. Sentence length is for source sentences.

LSTM	Gaze	ZIFF-DAVIS	BROADCAST			GOOGLE
Baseline		0.5668	0.7386	0.7980	0.6802	0.7980
Multitask	FP	0.6416	0.7413	0.8050	0.6878	0.8028
	REGR.	0.7025	0.7368	0.7979	0.6708	0.8016
Cascaded	FP	0.6732	**0.7519**	0.8189	**0.7012**	**0.8097**
	REGR.	**0.7418**	0.7477	**0.8217**	0.6944	0.8048

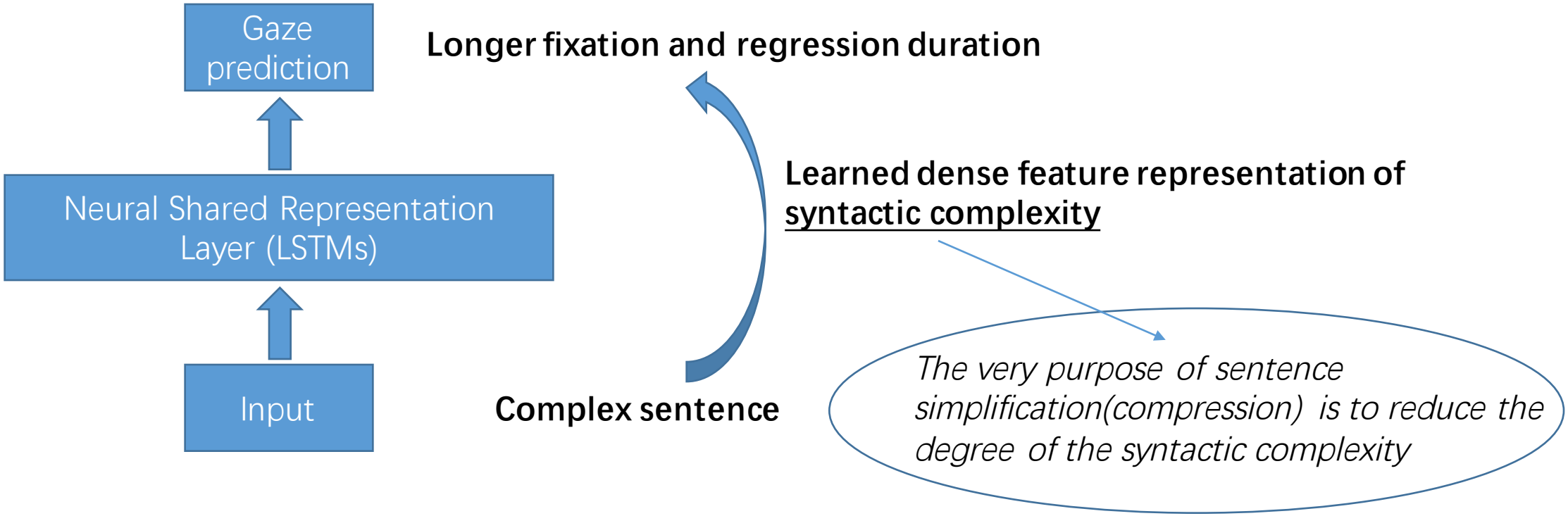
Table 3: Results (F_1). For all three datasets, the inclusion of gaze measures (first pass duration (FP) and regression duration (Regr.)) leads to improvements over the baseline. All models include CCG-supertagging as an auxiliary task. Note that BROADCASTwas annotated by three annotators. The three columns are, from left to right, results on annotators 1–3.



- Improvements using both first pass duration and regression duration (associated with interpretation of content).

Conclusion

- Revisiting the definition of MTL: main tasks use the [domain-specific information] of [related tasks] as a inductive bias to improve the generalization performance of the main task.



A Multi-task Learning Approach for Improving Product Title Compression with User Search Log Data

Jingang Wang, Junfeng Tian, Long Qiu, Sheng Li , Jun Lang, Luo Si and Man Lan

AAAI 2018

https://arxiv.org/pdf/1801.01725.pdf

Background – Compressing the lengthy titles on C2C online shopping platform

(a) *Search Result Page* (b) *Product Detail Page*

Figure 1: When a user issues a query "floral-dress long-sleeve women", the complete title cannot be displayed in the Search Result Page, unless the user proceeds to the detail page further.

- The title need to be compressed to be display on screens with small size.
- Problem: how to compress while retaining the frequent words in user log data

Example for original title and user search queries

Original Title	D'ZZIT 地素 秋 专柜 新款 丝绒 拉链 设计 半身 短裙 (D'ZZIT DiSu Autumn Counter New Silk Zipper Designed Half-length Skirt)	MIUCO 女装 夏季 新款 金线 刺绣 高腰 A字 摆 牛仔 背带 连衣裙 (MIUCO Women Summer New Gold-thread Embroidery High-waist A-line Jeans Braces Dress)
Top User Search Queries	D'ZZIT 短裙 丝绒 短裙 D'ZZIT 丝绒 短裙 ...	牛仔裙 连衣裙 牛仔 连衣裙 牛仔裙 ...

Figure 2: A triplet example. Both the compressed title and the query can help recognize important information from the original title.

Model

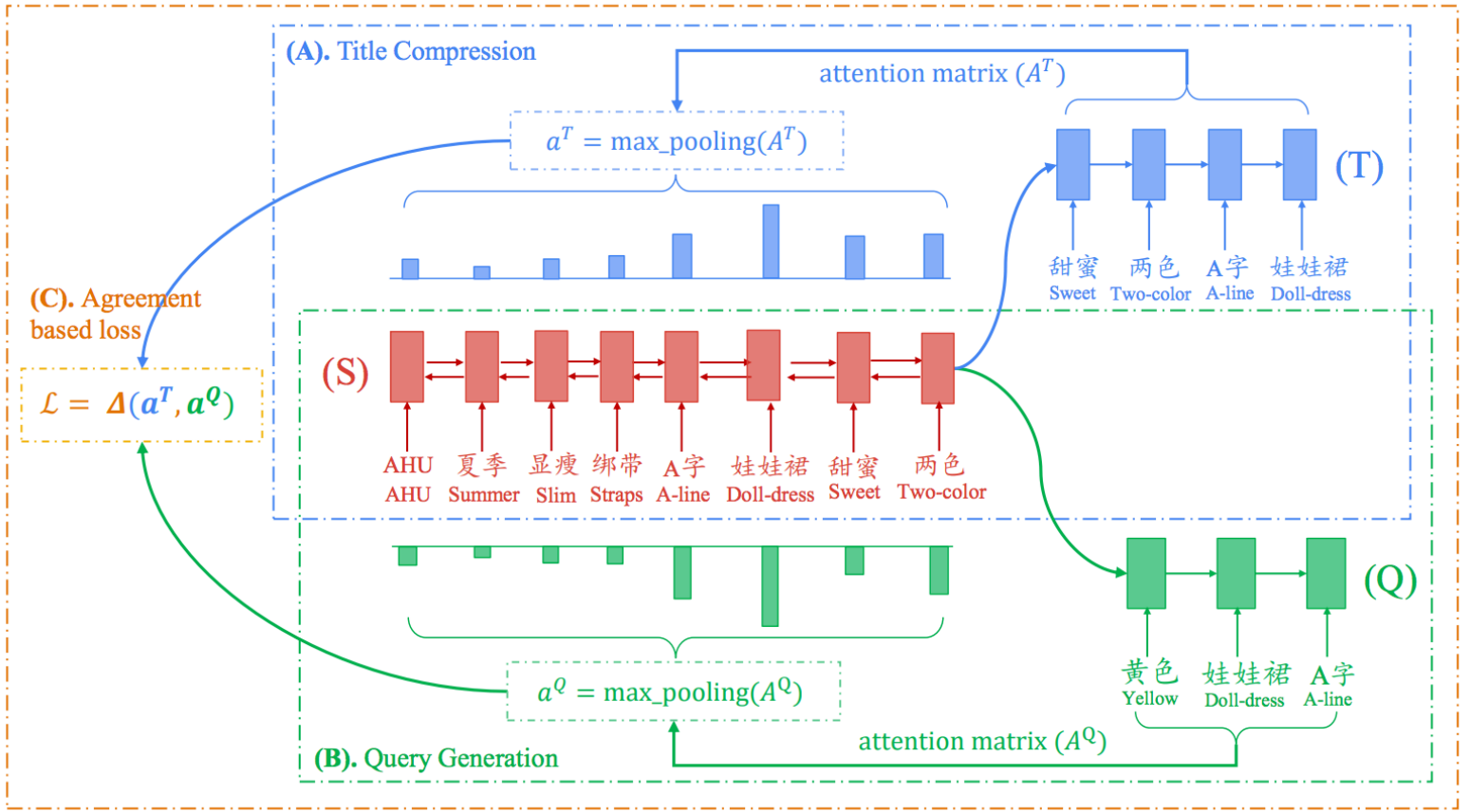


Figure 3: Multi-task Learning Framework, including two seq2seq components sharing the identical encoder. The main task is a Pointer Network to automatically point (select) the most informative words as compressed title. The auxiliary task is a standard seq2seq model to generate user search query. We utilize the attention distribution generated from user query to encourage the main task to agree on identity words.

Results

Table 2: ROUGE performance of various methods on the test set.

Method	ROUGE-1	ROUGE-2	ROUGE-L
Trunc.	30.43	19.13	29.00
ILP	48.28	29.84	43.65
Ptr-Net	69.03	55.30	67.98
Vanilla-MTL	65.92	52.94	65.20
Agree-MTL	**70.89**	**56.80**	**69.61**

Table 3: Manual evaluation results, including average core product recognition accuracy (Avg. Accu), average readability score (Avg. Read) and average informativeness score (Avg. Info).

Method	Avg. Accu	Avg. Read	Avg. Info
Trunc.	8.33 %	1.93	1.96
ILP	93.33%	4.63	3.90
Ptr-Net	**98.33 %**	4.66	4.13
Vanilla-MTL	96.67%	4.63	3.90
Agree-MTL	**98.33 %**	**4.80**	**4.66**

Vanilla-MTL $\mathcal{L} = \lambda \mathcal{L}_T + (1 - \lambda)\mathcal{L}_Q$

Agree-MTL $\mathcal{L} = \lambda_1 \mathcal{L}_T + \lambda_2 \mathcal{L}_Q + (1 - \lambda_1 - \lambda_2)\mathcal{L}_{agree}$

$$\mathcal{L}_{agree} = KL\left(a^T \parallel a^Q\right)$$

Conclusion

1. Identify related tasks
   - eye-movement information is a predictor of syntactic complexity.
   - User search query is indicative of which word should be retained.

2. They are not explicit features. Instead, they need to be implicitly extracted through neural layer.

3. The way of taking advantage of the information is also important, e.g. $\mathcal{L}_{agree} = KL\left(a^T \parallel a^Q\right)$

Thank you