Matching the Blanks: Distributional Similarity for Relation Learning

Livio Baldini Soares **Nicholas FitzGerald** **Jeffrey Ling*** **Tom Kwiatkowski**
Google Research
{liviobs,nfitz,jeffreyling,tomkwiat}@google.com

First: Better way to use BERT to extract relation

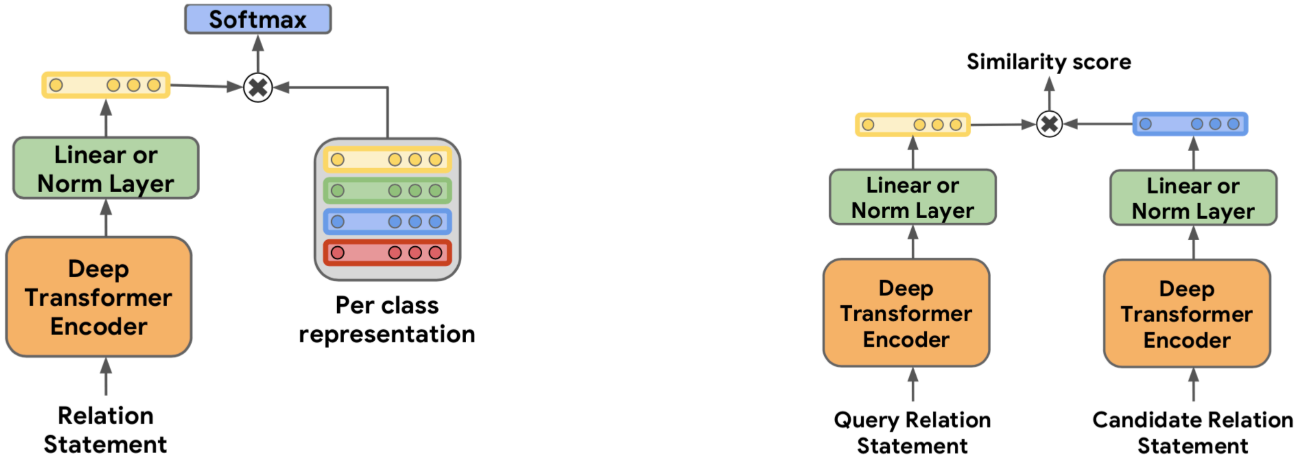
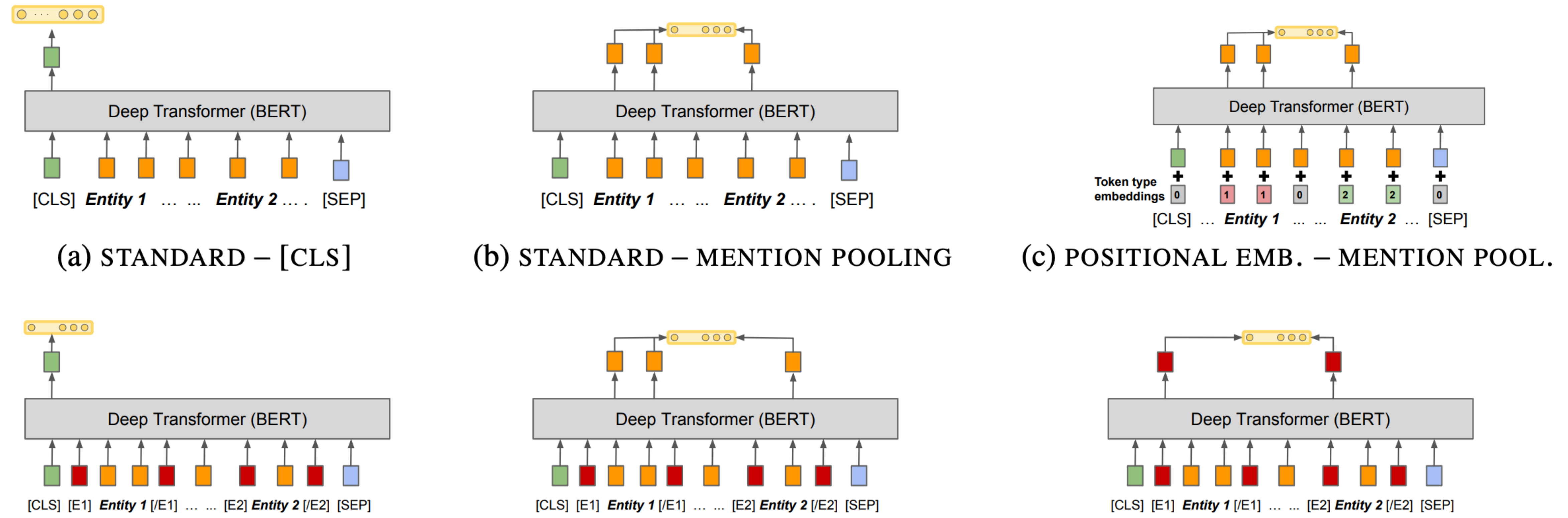


Figure 2: Illustration of losses used in our models. The left figure depicts a model suitable for supervised training, where the model is expected to classify over a predefined dictionary of relation types. The figure on the right depicts a pairwise similarity loss used for few-shot classification task.

Supporting Set	
(A) capital_of	(1) *London* is the capital of *the U.K.* (2) *Washington* is the capital of *the U.S.A.*
(B) member_of	(1) *Newton* served as the president of *the Royal Society*. (2) *Leibniz* was a member of *the Prussian Academy of Sciences*.
(C) birth_name	(1) *Samuel Langhorne Clemens*, better known by his pen name *Mark Twain*, was an American writer. (2) *Alexei Maximovich Peshkov*, primarily known as *Maxim Gorky*, was a Russian and Soviet writer.
Test Instance	
(A) or (B) or (C)	*Euler* was elected a foreign member of *the Royal Swedish Academy of Sciences*.

Table 1: An example for a 3 way 2 shot scenario. Different colors indicate different entities, blue for head entity, and red for tail entity.



(a) STANDARD – [CLS]
(b) STANDARD – MENTION POOLING
(c) POSITIONAL EMB. – MENTION POOL.
(d) ENTITY MARKERS – [CLS]
(e) ENTITY MARKERS – MENTION POOL.
(f) ENTITY MARKERS – ENTITY START

		SemEval 2010 Task 8		KBP37		TACRED		FewRel 5-way-1-shot
# training annotated examples		8,000 (6,500 for dev)		15,916		68,120		44,800
# relation types		19		37		42		100
		Dev F1	Test F1	Dev F1	Test F1	Dev F1	Test F1	Dev Acc.
Wang et al. (2016)*		–	88.0	–	–	–	–	–
Zhang and Wang (2015)*		–	79.6	–	58.8	–	–	–
Bilan and Roth (2018)*		–	84.8	–	–	–	68.2	–
Han et al. (2018)		–	–	–	–	–	–	71.6
Input type	**Output type**							
STANDARD	[CLS]	71.6	–	41.3	–	23.4	–	85.2
STANDARD	MENTION POOL.	78.8	–	48.3	–	66.7	–	87.5
POSITIONAL EMB.	MENTION POOL.	79.1	–	32.5	–	63.9	–	87.5
ENTITY MARKERS	[CLS]	81.2	–	68.7	–	65.7	–	85.2
ENTITY MARKERS	MENTION POOL.	80.4	–	68.2	–	69.5	–	87.6
ENTITY MARKERS	ENTITY START	**82.1**	**89.2**	**70**	**68.3**	**70.1**	**70.1**	**88.9**

Second: Match the blank. Unsupervised pre-training for relation extraction.

declare that for any pair of relation statements $\mathbf{r}$ and $\mathbf{r}'$, the inner product $f_\theta(\mathbf{r})^\top f_\theta(\mathbf{r}')$ should be high if the two relation statements, $\mathbf{r}$ and $\mathbf{r}'$, express semantically similar relations. And, this inner product should be low if the two relation statements express semantically different relations.

$$p(l=1|\mathbf{r},\mathbf{r}') = \frac{1}{1+\exp f_\theta(\mathbf{r})^\top f_\theta(\mathbf{r}')}$$

$$\begin{aligned}\mathcal{L}(\mathcal{D}) = -\frac{1}{|\mathcal{D}|^2} \sum_{(\mathbf{r},e_1,e_2)\in\mathcal{D}} \sum_{(\mathbf{r}',e_1',e_2')\in\mathcal{D}} \\ \delta_{e_1,e_1'}\delta_{e_2,e_2'} \cdot \log p(l=1|\mathbf{r},\mathbf{r}') + \\ (1-\delta_{e_1,e_1'}\delta_{e_2,e_2'}) \cdot \log(1-p(l=1|\mathbf{r},\mathbf{r}'))\end{aligned} \quad (1)$$

$\mathbf{r}_A$	In 1976, $\boldsymbol{e_1}$ (then of Bell Labs) published $\boldsymbol{e_2}$, the first of his books on programming inspired by the Unix operating system.
$\mathbf{r}_B$	The "$\boldsymbol{e_2}$" series spread the essence of "C/Unix thinking" with makeovers for Fortran and Pascal. $\boldsymbol{e_1}$'s Ratfor was eventually put in the public domain.
$\mathbf{r}_C$	$\boldsymbol{e_1}$ worked at Bell Labs alongside $\boldsymbol{e_3}$ creators Ken Thompson and Dennis Ritchie.
Mentions	$\boldsymbol{e_1}$ = Brian Kernighan, $\boldsymbol{e_2}$ = Software Tools, $\boldsymbol{e_3}$ = Unix

70%: replace the mention with [BLANK]

	SemEval 2010	**KBP37**	**TACRED**
SOTA	84.8	58.8	68.2
BERT_{EM}	89.2	68.3	70.1
BERT_{EM}+MTB	**89.5**	**69.3**	**71.5**

	5-way 1-shot	**5-way 5-shot**	**10-way 1-shot**	**10-way 5-shot**
Proto Net	69.2	84.79	56.44	75.55
BERT_{EM}+MTB	**93.9**	**97.1**	**89.2**	**94.3**
Human	92.22	–	85.88	–

More Data, More Relations, More Context and More Openness: A Review and Outlook for Relation Extraction

Xu Han[1*], Tianyu Gao[1*], Yankai Lin[2*], Hao Peng[1], Yaoliang Yang[1], Chaojun Xiao[1], Zhiyuan Liu[1], Peng Li[2], Maosong Sun[1], Jie Zhou[2]
[1] State Key Lab on Intelligent Technology and Systems,
Institute for Artificial Intelligence,
Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China

Benchmark	Normal	ME	OE
Wiki80 (Acc)	0.861	0.631	0.763
TACRED (F-1)	0.666	0.211	0.412
NYT-10 (AUC)	0.349	0.216	0.185
Wiki-Distant (AUC)	0.222	0.145	0.173

The observation is contrary to human intuition: we classify the relations between given entities mainly from the text description, yet models learn more from their names. To make real progress in understanding how language expresses relational facts, this problem should be further investigated and more efforts are needed.

Our experiment: mask mention to avoid overfitting => not work

Document-level Representation Learning using Citation-informed Transformers

Arman Cohan[†*] **Sergey Feldman**[†*] **Iz Beltagy**[†] **Doug Downey**[†] **Daniel S. Weld**[†,‡]

[†]Allen Institute for Artificial Intelligence

[‡]Paul G. Allen School of Computer Science & Engineering, University of Washington

{armanc,sergey,beltagy,dougd,danw}@allenai.org

Drawback of BERT => domain-specific pretrain => strong assumption

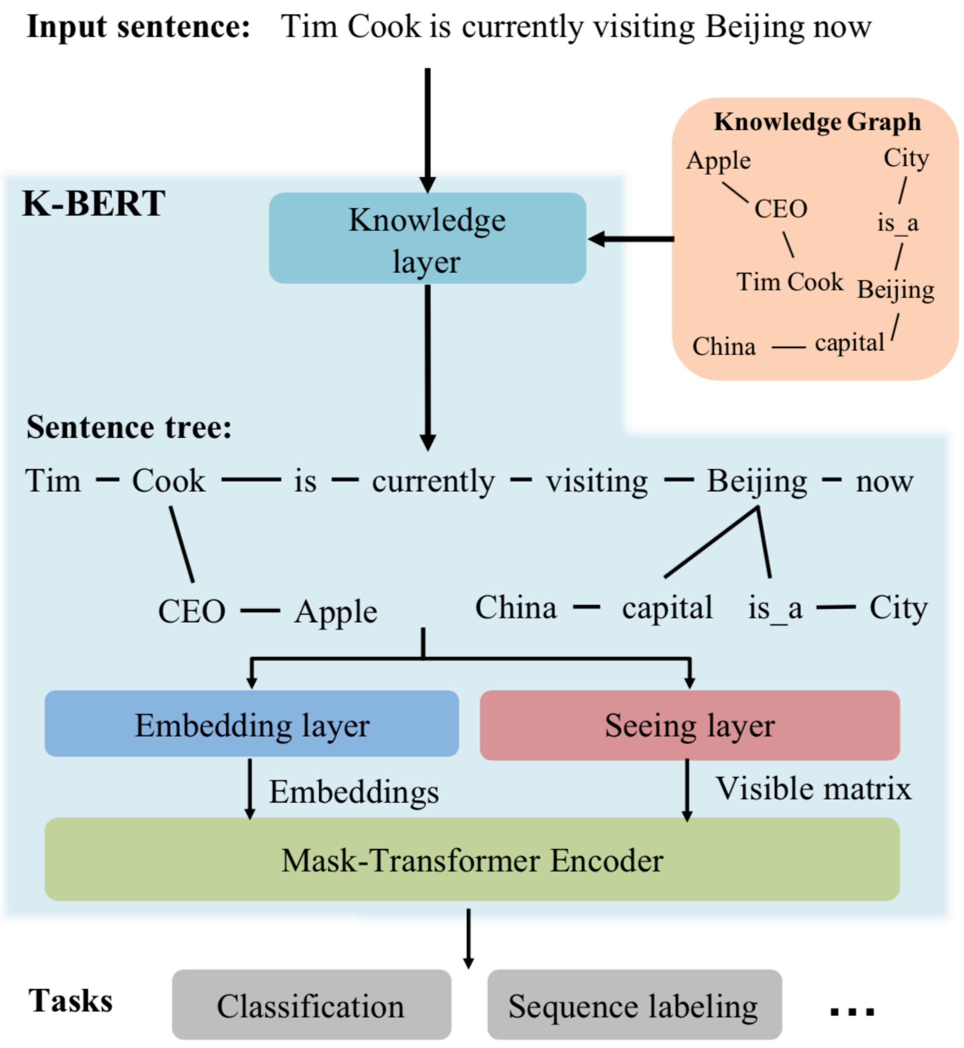
$$\mathcal{L} = \max\left\{\left(\mathrm{d}\left(\mathcal{P}^Q, \mathcal{P}^+\right) - \mathrm{d}\left(\mathcal{P}^Q, \mathcal{P}^-\right) + m\right), 0\right\}$$

K-BERT: Enabling Language Representation with Knowledge Graph

Weijie Liu[1], Peng Zhou[2], Zhe Zhao[2], Zhiruo Wang[3], Qi Ju[2,*], Haotang Deng[2] and Ping Wang[1,*]
[1]Peking University, Beijing, China
[2]Tencent Research, Beijing, China
[3]Beijing Normal University, Beijing, China

1. Heterogeneous Embedding Space
2. Knowledge Noise

Inject the knowledge from KG into BERT



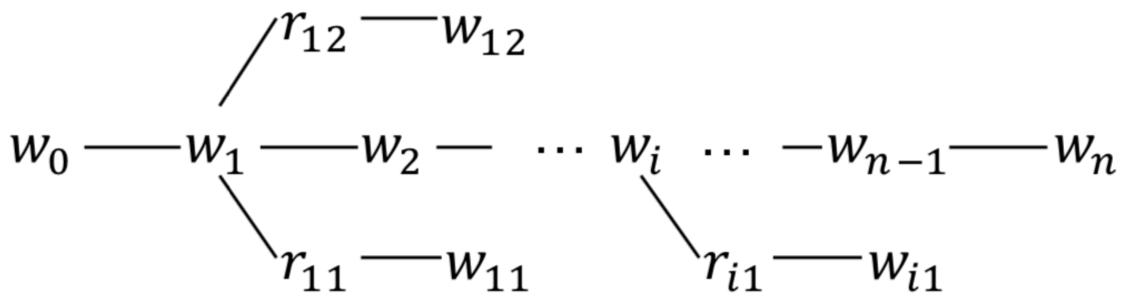
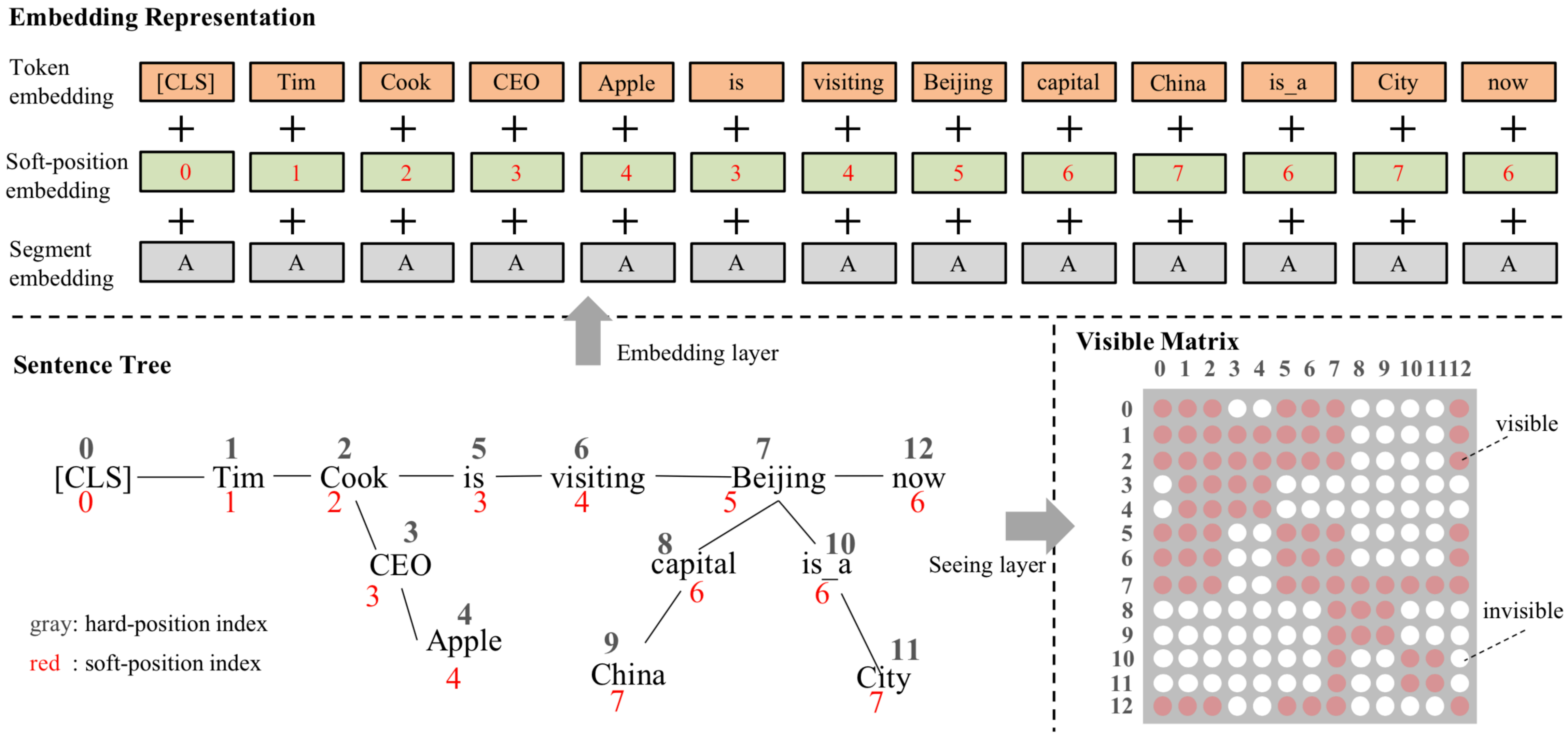


Figure 3: Structure of the sentence tree.



Embedding Representation
Token embedding
[CLS]
Tim
Cook
CEO
Apple
is
visiting
Beijing
capital
China
is_a
City
now
Soft-position embedding
0 1 2 3 4 3 4 5 6 7 6 7 6
Segment embedding
A A A A A A A A A A A A A
Embedding layer
Sentence Tree
[CLS]
Tim
Cook
is
visiting
Beijing
now
CEO
Apple
capital
is_a
China
City
gray: hard-position index
red : soft-position index
Seeing layer
Visible Matrix
0 1 2 3 4 5 6 7 8 9 10 11 12
visible
invisible

Table 1: Results of various models on sentence classification tasks on open-domain tasks ($Acc.$ %)

Models\Datasets	Book_review		Chnsenticorp		Shopping		Weibo		XNLI		LCQMC	
	Dev	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*
Pre-trainied on WikiZh by Google.												
Google BERT	88.3	**87.5**	93.3	94.3	96.7	96.3	98.2	98.3	76.0	75.4	88.4	86.2
K-BERT (HowNet)	88.6	87.2	**94.6**	**95.6**	**97.1**	**97.0**	98.3	98.3	**76.8**	**76.1**	**88.9**	86.9
K-BERT (CN-DBpedia)	**88.6**	87.3	93.9	95.3	96.6	96.5	98.3	98.3	76.5	76.0	88.6	**87.0**
Pre-trained on WikiZh and WebtextZh by us.												
Our BERT	**88.6**	87.9	94.8	95.7	96.9	**97.1**	98.2	98.2	77.0	76.3	89.0	86.7
K-BERT (HowNet)	88.5	87.4	**95.4**	95.6	96.9	96.9	98.3	**98.4**	**77.2**	**77.0**	**89.2**	**87.1**
K-BERT (CN-DBpedia)	88.8	87.9	95.0	**95.8**	**97.1**	97.0	98.3	98.3	76.2	75.9	89.0	86.9

Table 2: Results of various models on NLPCC-DBQA (MRR %) and MSRA-NER ($F1$ %).

Models\Datasets	NLPCC-DBQA		MSRA-NER	
	Dev	*Test*	*Dev*	*Test*
Pre-trained on WikiZh by Google.				
Google BERT	93.4	93.3	94.5	93.6
K-BERT (HowNet)	93.2	93.1	95.8	94.5
K-BERT (CN-DBpedia)	**94.5**	**94.3**	**96.6**	**95.7**
Pre-trained on WikiZh and WebtextZh by us.				
Our BERT	93.3	93.6	95.7	94.6
K-BERT (HowNet)	93.2	93.1	96.3	95.6
K-BERT (CN-DBpedia)	**93.6**	**94.2**	**96.4**	**95.6**

Table 3: Results of various models on specific-domain tasks (%).

Models\Datasets	Finance_Q&A			Law_Q&A			Finance_NER			Medicine_NER		
	P.	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*	*P.*	*R.*	*F1*
Pre-trained on WikiZh by Google.												
Google BERT	81.9	86.0	83.9	83.1	90.1	86.4	84.8	87.4	86.1	91.9	93.1	92.5
K-BERT (HowNet)	83.3	84.4	83.9	83.7	91.2	87.3	86.3	89.0	**87.6**	93.2	93.3	93.3
K-BERT (CN-DBpedia)	81.5	88.6	**84.9**	82.1	93.8	**87.5**	86.1	88.7	87.4	93.9	93.8	93.8
K-BERT (MedicalKG)	-	-	-	-	-	-	-	-	-	94.0	94.4	**94.2**
Pre-trained on WikiZh and WebtextZh by us.												
Our BERT	82.1	86.5	84.2	83.2	91.7	87.2	84.9	87.4	86.1	91.8	93.5	92.7
K-BERT (HowNet)	82.8	85.8	84.3	83.0	92.4	87.5	86.3	88.5	87.3	93.5	93.8	93.7
K-BERT (CN-DBpedia)	81.9	87.1	**84.4**	83.1	92.6	**87.6**	86.3	88.6	**87.4**	93.9	94.3	94.1
K-BERT (MedicalKG)	-	-	-	-	-	-	-	-	-	94.1	94.3	**94.2**

1. 跑实验，写paper，进度40%
2. 在CDR数据集上尝试，效果不佳。猜测原因可能是 a) 数据集规模较小， train/dev/test各500个doc；b）类别简单，只有Chemical和Disease，且关系类型只有一种 就是Chemical->Disease; c) 生化领域

3. 衡量模型inter-sentence information aggregation 和 relation inference能力。

⇒ Subset of DocRED dev set

⇒ evaluation metric?

⇒ doc -> relation labels

⇒ 1. inter-sentence relation; 2. intra-sentence relation

⇒ Recall容易计算，模型预测出来属于1的inter-sentence relation / # inter-sent relation
⇒ how to define precison? 模型预测出来的relation中，可能有inter-sent的，也有intra-sent的，还有无中生有的，如何计算presion呢？

⇒ 其他指标？

4. Model: Encode/mention-level graph Aggregation/entity-level graph Inference/Classfify

GAIN: Graph Aggregation & Inference Network for Document-level Relation Extraction

No pain, No GAIN: ~