Extracting Zero-shot Structured Information from Form-like Documents: Pretraining with Keys and Triggers

Rongyu Cao, Ping Luo



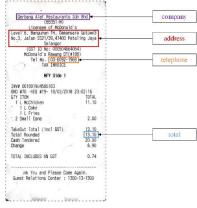
Contents

- Background
- Challenge
- Solution
- Experiments

Background

Information Extraction from Document

- Input
 - a document (a collection of words with its content and position).
 - given attributes (keys).
- Output
 - for each key, which words consist the corresponding value.



Information Extraction from Plain Text

• Input

V.

S.

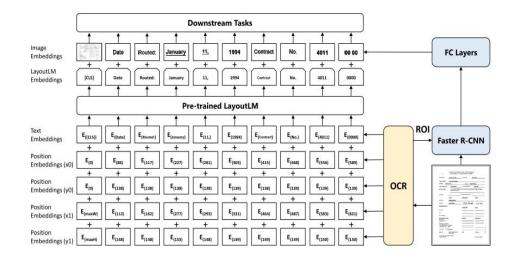
- a sentence (a sequence of words with its content).
- given attributes (keys).
- Output
 - for each key, which words consist the corresponding value.

Evkoodance New Style Khaki NUbuck Size US 4-12 With Rhinestone High Quality 8.5cm Heel Open Toe Latin Dance Shoes

Attributes: color, model number, athletic shoe type, heel height

Background

- Previous Method: LayoutLM [1]
- Framework
 - Like BERT, take a collection of words with text embeddings and 2d position embeddings as input.
 - 2. Like BERT, use multi-layers, multiheads transformer.
 - 3. Predict a class label (multiclassification) for each word.
 - 4. The parameters are pre-trained on two tasks: masked visuallanguage model and multi-label document classification.

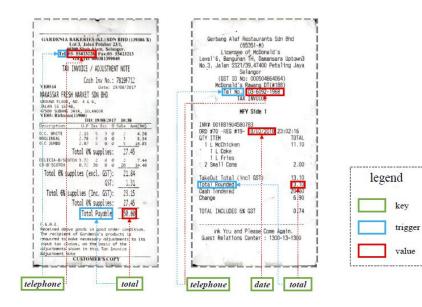


Challenge

- In real-world production, the keys needed to extract might have a huge number.
- Labeling large-scale training data for each key is costly.
- So, can the model generalize to unseen keys (zero-shot keys) without additional annotation?
- Thus, we should let the model learn key-invariant feature representation.

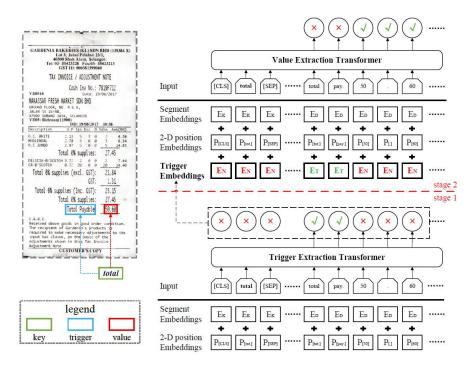
Solution

- What is the key-invariant feature representation?
- No matter what is the given key, the way we find value for the key is the same.
- Which way? How a human find the value for a given key?



Solution

- Model: Two-stages LayoutLM-based model.
- It explicitly learns two mappings: key-to-trigger and trigger-to-value.



Solution

 How to annotate triggers and values for large-scale documents to pre-train this model?

Artis	e Talk	Read Edit Verwhistory	Search whopeda	3
J	ay Chou			
Fr	om Wikipedia, the free encyclopedia			
	In this Chinese name, the family name is Chou (III).			
		see (Emildle pinylin: Zhóu Jiélún; born 18 January 1979) ^[1] is a Taiwanese singer, songwriter, rapper, record producer, actor, film director, businessman and magician. Dubbed the "King of Mandopop", and having sold more than 30 million albums, Chou is one of the best-selling artists in Mainland	1	Jay Chou
		ing, who he has frequently collaborated with on his music.	4 <u> </u>	周杰倫
		00, under the record company Alfa Mukit to moderate success. Chour ones to fame with the release of his second studio album, <i>Fastage</i> (2010), which combined Western and Eastern music styles. The album von five Golden Melody Awards, including Album of the Year. He has since further of his integrad and printing significant proceedimenters in a Alass communities around the world to come frant or multice styles. The album von five Golden Melody Awards, including Album of the Year. He has since further of his integrad and printing significant proceediments in a start world. For has area marked on six world to vori pto smoot the vori of to note than 10 million partetors as of 2010. ^[2]	() ()	
		ment company JVR Music. ^[3] Outside of music, Chou has served as the President of his own fashion brand PHANTACI since 2006. As an actor, Chou made his acting debut in the film <i>Initial D</i> (2005), followed shortly by a starring role in the epic <i>Curse of the Golden Flower</i> (2008). He has since starr	rd .	-
in	a number of movies, becoming known to Western a	udiences when he made his Hollywood debut in 2011 with The Green Hornet, starring alongside Seth Rogen and Christoph Waltz, followed by Now You See Me 2 (2016).		
	Contents [hide]			ALLA
	Early life		4	
	Early career			
	Music career			12300
	3.1 2000 to 2003 3.2 2003 to 2005		4	Ex H
	3.3 Single released in 2019			104201
	3.4 Single released in 2019 3.4 Single released in 2020		·	199
	Musical style			
1	4.1 Lyrics			- HIN CON
	4.2 Chinese cultural elements		1	
	Collaborations			Chou in 2016
	World tours		Born	18 January 1 Linkou, Taip
	Movie career		1	Taiwan
	7.1 Acting		Nationality	Republ
	7.2 Directing		Other names	President C
	Other works		Occupation	Singer • son
	8.1 Book: Grandeur de D Major			producer • r
	8.2 Endorsements		4	film director
	8.3 Philanthropy			personality
	8.4 J Gaming		Years active	2000-prese
5	Media and public relations		Spouse(s)	Hannah Qu
	9.1 Public image		Children	2
	9.2 Response to the news media and paparazzi		Parents	Chou Yao-C
	9.3 Fanbase		Awards	Yeh Hui-Me Full list
	0 Personal life 10.1 Marriage		Awards	Musical career
	10.1 Marriage 10.2 Religious beliefs		Genres	Mandopop
	1 Discography			Latin soul
	2 Filmography		<u>.</u>	hop
	12.1 Film		Instruments	Vocals • pia
	12.2 Television series			guitar • guz flute • drum
	12.3 Television shows		Labels	Sony Music
	12.3.1 Adapted musical			present)
	3 Accolades			JVR Music (
	4 Notes		1	Alfa Music (
	5 References			Chinese name
	6 External links		Traditional Chi	
			Simplified Chir Transcriptions	
	arly life (edit)			·

Chou was conscripted for mandatory military service after graduating from high school with inadequate grades for university. However, severe back pain triggered by sports eventually led to the diagnosis of ankylosing spondylitis (HLA-B27), and he was exempted from military service.

[2] Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; et al. DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic web. 2015.

Experiments

• Datasets

	description	acces s	#key	#zero- shot	#document	#pair
En- Infobox	Infobox on English Wikipedia webpage	crawle r	7,370	-	398,467	1,238,281
Zh- Infobox	Infobox on Chinese Wikipedia webpage	crawle r	6,151	-	281,281	897,520
SROIE	English receipts	public	6	2	972	5,505
Grater	Chinese bank statements	private	24	8	4,032	18,825

Zero-shot keys for SROIE dataset: cash, telephone.

Zero-shot keys for Grater dataset: organization (开户机构), balance in previous page (上页余额), subject (科目), card number (卡号), balance (本页余额), unit (单位), print method (打印渠道), credit count (信用笔数).

Experiments

Results

row		the SROIE dataset						the Grater dataset					
	method	zero-shot keys			non-zero-shot keys			zero-shot keys			non-zero-shot keys		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	BERT [Devlin et al. 2019]	0	0	0	0.9099	0.9099	0.9099	0	0	0	0.9986	0.9965	0.9976
2	LayoutLM [Xu et al. 2020b]	0	0	0	0.9460	0.9460	0.9460	0	0	0	1.0000	1.0000	1.0000
3	KA-bert	0.1043	0.1000	0.1021	0.9373	0.9373	0.9373	0.2431	0.2422	0.2426	0.9993	0.9979	0.9986
4	KA-wiki	0.6468	0.5961	0.6204	0.9469	0.9430	0.9449	0.3768	0.3702	0.3735	1.0000	1.0000	1.0000
5	KATA-bert	0.0649	0.0471	0.0545	0.9366	0.9382	0.9347	0.3692	0.2491	0.2975	0.9993	0.9993	0.9993
6	KATA-wiki	0.7678	0.6549	0.7069	0.9483	0.9457	0.9470	0.7378	0.7301	0.7339	1.0000	1.0000	1.0000
7	KATA-wiki w/ index embedding	0.6569	0.6000	0.6283	0.9455	0.9423	0.9439	0.6809	0.6644	0.6725	1.0000	1.0000	1.0000
8	KATA-wiki w/o fine-tuning	0.5116	0.4314	0.4681	0.1106	0.0502	0.0691	0.1322	0.0796	0.0994	0.0963	0.0594	0.0735

Table 2: Comparing KATA with baseline models.

Conclusion

- We propose a novel task to extract the values of zero-shot keys in documents.
- We propose a two-stages framework to learn key-to-trigger and trigger-to-value mappings.
- We construct large-scale labeled dataset to pre-train the model on Wikipedia Infobox.

国际人工智能会议 AAAI 2021论文北京预讲会

THANKS

2020.12.19