Contextual Word Embeddings

2020年6月26日 14:52

##00_背景

目前通过word2vec, GloVe和fasttext可以获取词向量，但是存在两个问题：
① 不考虑词出现的上下文，一个词类将会是同一个表示；而需要一个更加细粒度的无歧义的词义；
② 一个词的表示，而词有很多不同的方面，包括语义、语法行为和语体风格、隐含意义

##01_相关工作

- 预训练词向量：word2vec, GloVe, fasttext
  - 从随机词向量开始效果不好
  - 使用预训练词向量能带来好处
- POS（Part-of-Speed tagging，词性标注）确定每个词是名词、动词、形容词或其他词性
- NER（Named Entity Recognition，命名实体识别）指识别文本中具有特定意义的实体，包括人名、地名、机构名、专有名词等；
- Syntax Parsing（句法解析），包括句法结构分析（Syntax structure parsing）和依存关系分析（denpdency parsing）
- 预训练语言模型

##02_模型

- ELMo
- UMLfit->GPT(Generative Pre-trained Transformer, OpenAI)
- NER（Named Entity Recognition）
- Transformer
  - Attention
    - Dot-Production Attention
    - Scaled dot-production attention
    - Self-attention in the encoder
    - Multi-head attention
- Bert(Bidirectional Encoder Representation from Transformers)

##03_实验

- GLUE 任务
- NER任务：TagLM
- NMT任务：CoVe

##04_总结

参考文献

Smith N A . Contextual Word Representations: A Contextual Introduction[J]. 2019.	Context Word...	
http://jalammar.github.io/illustrated-bert/		
Peters M E , Neumann M , Iyyer M , et al. Deep contextualized word representations[J]. 2018.	Deep contextu...	
https://zhuanlan.zhihu.com/p/51679783	ELMo	
Collobert R , Weston J , Bottou, Léon, et al. Natural Language Processing (almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.	Natural Languag...	
Collobert R , Weston J . A unified architecture for natural language processing: Deep neural networks with multitask learning[C]// Machine Learning, Proceedings of the		

Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008. 2008.	A Unified Architect...	
http://www.dataguru.cn/article-14265-1.html	一文看懂NLP神经网络发展历史中最重要的8个里程碑!	

Mikolov T . Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.	Distribut... Represe...	• Continous Skip-gram model
		• Speedup
		  • subsampling of the frequent words
		  • Hierarchical softmax called negative sampling

	Semi-supervis...	•
Mccann B , Bradbury J , Xiong C , et al. Learned in Translation: Contextualized Word Vectors[J]. 2017.	Learned in Translati...	
Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.	Attention Is All Yo...	• Transformer based solely on attention mechanism, dispensing with recurrence and convolutions entirely • Follows this overall architecture using stacked self-attenttion and point-wise, fullly connected layers for both the encoder and decoder • Self-attention called intra-attention is an attention mechanism relating differnet positions of a single sequence in order to compute a representation of the sequence • Attention function can be described as mapping a query and a set of key-value pairs to an output, where the query ,keys, values and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key
Cho K , Van Merrienboer B , Gulcehre C , et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer ence, 2014.	Learning Phrase...	
Collobert R , Weston J , Bottou, Léon, et al. Natural Language Processing (almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.	Natural Languag...	
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	BERT Pre-training...	