选题方向整理

1. Real Multi-Sense or Pseudo Multi-Sense: An Approach to Improve Word Representation

<https://arxiv.org/abs/1701.01574>

word2vec提供的词向量通常有许多问题，比如一个典型的问题是一词多义，于是很多的工作都是在解决一词多义的问题，但一个词对应的多个向量其实都指向同一个词义，本文的工作正是对这些伪一词多义进行识别，降低语言研究的复杂度。

1. Multi-level Representations for Fine-Grained Typing of Knowledge Base Entities

<https://arxiv.org/abs/1701.02025>  
<http://cistern.cis.lmu.de/figment/>

entity是知识图谱的基础组件，很多的entity都是罕见词（短语），entity的表示是一个相对困难的问题。本文提出了一种char-level、word-level和entity-level三种level的联合表示模型，得到了不错的效果。

1. NLP Service APIs and Models for Efficient Registration of New Clients

<https://arxiv.org/abs/2010.01526>

最先进的NLP推理使用了大量的神经架构和经过gpu数月训练的模型，远远超出了大多数NLP用户的能力范围。这导致了大型人工智能公司的通用的基于api的公共NLP服务模式，为大量客户提供服务。无论是(硬件不足的)客户端还是(大量订阅的)服务器都无法负担传统的微调。许多客户拥有很少或没有标记的数据。我们开始研究如何将集中式NLP服务适应于客户，并提出了一种实用且轻量级的方法。每个客户机使用无监督的、基于语料库的草图注册到服务。服务器使用一个辅助网络将草图映射到一个抽象的向量表示，然后通知主标注网络。当一个新客户用它的草图注册时，它可以立即获得准确性的好处。我们使用情感标记、NER和预测语言建模来证明所提出的架构是成功的。

1. MacBert：Revisiting Pre-trained Models for Chinese Natural Language Processing

<https://arxiv.org/pdf/2004.13922.pdf>

[hfl/chinese-macbert-base](https://huggingface.co/hfl/chinese-macbert-base) ; [hfl/chinese-macbert-large](https://huggingface.co/hfl/chinese-macbert-large)

作者对原有的BERT的MLM任务进行了提升，不使用[MASK]作为掩码，而是使用相似的字进行掩码，然后发现Performance提升了，起个新名字MacBERT。提出了新的MacBert模型，其缓和了pre-training阶段和fine-tuning阶段的gap。采用的方式是“mask字时，采用相似的字进行mask”。