融合语言知识的神经网络中文词义消歧模型

穆玲玲，　程晓煜，　咎红英，　韩英杰

(郑州大学 信息工程学院　河南 郑州 450001)

摘要：提出一种融合语言知识的神经网络中文词义消歧模型,在双向长短时记忆网络中使用目标词的释义和例句信息进行消歧。该模型在 SemEval-2007 中英文词义消歧数据集上的实验结果表明,融合语言知识后,词义消歧的宏平均准确率和微平均准确率分别比基线模型提高了 2.31%和 1.93%,说明在神经网络模型中融合语言知识有助于改善中文词义消歧的效果。

关键词：词义消歧；词典；神经网络；语言资源

中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1671-6841(2020)03-0015-06

DOI：10.13705/j.issn.1671-6841.2019384

0　引言

词义消歧是自然语言处理中的基础任务之一,用于确定目标词在特定上下文语境的词义[1],是信息抽取、机器翻译和阅读理解等任务的基础。词义消歧主要有 3 种方法:基于知识库的方法、有监督方法和无监督方法。其中有监督词义消歧通常使用传统机器学习模型实现,如支持向量机[2]、最大熵[3]和贝叶斯分类器[4]等,其准确率高于另外 2 种方法。

目前在有监督词义消歧任务中大量使用了神经网络模型[5],并取得了优于传统统计模型的结果。例如,文献[6-7]分别使用双向长短时记忆网络和多任务学习方法成功实现了词义消歧。基于神经网络的词义消歧方法虽然取得了较好的效果,但其存在以下两个问题:① 需要大规模的标注语料,否则将导致神经网络模型准确率下降。② 没有使用相关的语言知识,忽略了语言学家已建立的丰富资源。有研究表明,在神经网络中融合语言知识有助于提高模型的有效性,可以在保证准确率的前提下,降低模型训练对大规模标注语料的需求。文献[8]在循环神经网络中使用外部语言知识,提高了机器阅读的准确率。文献[9]在神经网络中使用了 WordNet 的释义信息,利用记忆网络[10-12]建模目标词上下文和释义的内在联系,在英文数据集上取得了非常高的准确率。文献[13-14]分别利用释义和 WordNet[15]中的语义增强了词义向量的表示,并将其作为 SVM 分类器的特征,使得词义消歧的准确率提高了 1%以上。文献[16]将 WordNet 的词根向量化后与 GloVe 词向量拼接,作为双向长短时记忆网络的输入用于词义消歧。上述研究均是针对英文词义消歧,而中文的神经网络词义消歧中融合语言知识的研究文献尚未被发现。本文在文献[9]基础上,利用外部记忆机制将目标词的释义和例句信息融入神经网络词义消歧模型中,通过注意力机制构建目标词的上下文与由释义和例句表示的词义之间的语义关系。在 SemEval-2007 中英文词义消歧数据集上的实验结果显示,本文模型的宏平均准确率和微平均准确率均比基线模型有所提高。

1　词义消歧模型

通过双向长短时记忆网络[17]分别实现目标词的上下文表示和目标词的词义表示,目标词词义由释义+例句联合表示,通过注意力机制构建目标词的上下文与词义之间的语义关系。

收稿日期:2019-08-23

基金项目:国家社会科学基金重大项目(18ZDA315);国家社会科学基金项目(17BXW065,14BYY096);河南省科技攻关项目(192102210260);河南省高等学校重点科研项目(20A520038)。

作者简介:穆玲玲(1969—),女,山东蓬莱人,副教授,主要从事自然语言处理研究,E-mail:iellmu@zzu.edu.cn。

融合释义、例句信息的词义消歧模型如图1所示,该模型包括上下文表示模块、词义表示模块、记忆模块和打分模块4个部分。

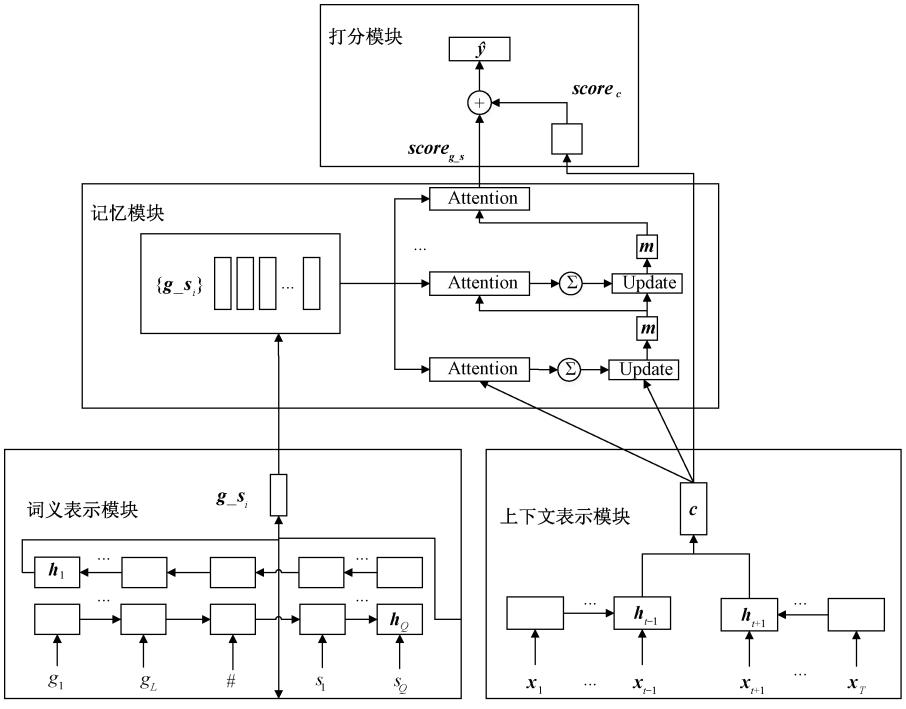

图1 融合释义、例句信息的词义消歧模型

Figure 1 A word sense disambiguation model leveraging glosses and example sentences

1.1 上下文表示模块

上下文表示模块采用双向长短时记忆网络构建目标词上下文的语义信息。设目标词 w_t 的上下文词语序列为 $[w_1,\cdots,w_{t-1},w_t,w_{t+1},\cdots,w_T]$($T$ 为上下文的长度),对应的词向量序列为 $[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{t-1},\boldsymbol{x}_t,\boldsymbol{x}_{t+1},\cdots,\boldsymbol{x}_T]$。将 $[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{t-1}]$ 作为前向长短时记忆网络的输入,得到在 $t-1$ 时刻的输出 $\overrightarrow{\boldsymbol{h}_{t-1}}$,将 $[\boldsymbol{x}_T,\boldsymbol{x}_{T-1},\cdots,\boldsymbol{x}_{t+1}]$ 作为后向长短时记忆网络的输入,得到在 $t+1$ 时刻的输出 $\overleftarrow{\boldsymbol{h}_{t+1}}$。将 $\overrightarrow{\boldsymbol{h}_{t-1}}$ 和 $\overleftarrow{\boldsymbol{h}_{t+1}}$ 拼接得到上下文向量 $\boldsymbol{c}$,用 $\boldsymbol{c}$ 来表示目标词上下文的语义信息。

1.2 词义表示模块

词义表示模块使用双向长短时记忆网络分别对释义和例句进行编码,作为目标词的词义表示向量。目标词 w 的第 i 个词义对应的释义和例句的词语序列分别记作 $\boldsymbol{g}=[g_1,g_2,\cdots,g_L]$ 和 $\boldsymbol{s}=[s_1,s_2,\cdots,s_Q]$,其中 L 和 Q 分别为释义和例句的词语长度。将释义和例句的词语序列拼接得到 $\boldsymbol{gs}=[g_1,g_2,\cdots,g_L,\#,s_1,s_2,\cdots,s_Q]$,其中"#"是释义和例句的区分符号。将 $\boldsymbol{gs}$ 转换为对应的词向量序列,作为双向长短时记忆网络的输入,分别得到前向和后向长短时记忆网络最后一个时刻的输出 $\overrightarrow{\boldsymbol{h}_Q}$ 和 $\overleftarrow{\boldsymbol{h}_1}$,拼接 $\overrightarrow{\boldsymbol{h}_Q}$ 和 $\overleftarrow{\boldsymbol{h}_1}$ 得到释义和例句联合表示向量 $\boldsymbol{g\_s}$,$\boldsymbol{g\_s}$ 即为目标词的词义表示。

1.3 记忆模块

记忆模块用于建模目标词的上下文向量与词义表示向量的语义关系,提取与上下文相关的词义信息,词义表示由释义、例句联合向量表示。该模块的输入为目标词的上下文向量 $\boldsymbol{c}$ 及其词义向量集合 $\{\boldsymbol{g\_s}_1,\boldsymbol{g\_s}_2,\cdots,\boldsymbol{g\_s}_N\}$($N$ 为目标词的词义个数),包括注意力计算和记忆向量更新2个部分,注意力计算建模上下文向量 $\boldsymbol{c}$ 与词义向量 $\boldsymbol{g\_s}$ 之间的语义信息。为了提高模型对释义、例句以及上下文语义的理解,记忆模块采用多轮注意力计算。在每轮计算后,根据当前计算结果更新记忆向量。

目标词第 i 个词义向量 $\boldsymbol{g\_s}_i$ 在第 k 轮的注意力 $\boldsymbol{a}_i^k$ 的计算方法可以表示为

$$\boldsymbol{a}_i^k=\frac{\exp(\boldsymbol{e}_i^k)}{\sum_{j=1}^{N}\exp(\boldsymbol{e}_i^j)},\ \boldsymbol{e}_i^k=\boldsymbol{g\_s}_i\cdot\boldsymbol{m}^{k-1},\tag{1}$$

式中:$\boldsymbol{m}^{k-1}$ 是第 $k-1$ 轮的记忆向量,初始记忆向量 $\boldsymbol{m}^0$ 使用上下文向量 $\boldsymbol{c}$。第一轮计算中,注意力反映的是词义向量和上下文向量的相似度,在以后每一轮的计算中,注意力反映的是词义向量与上一轮记忆向量的相似度。

为了突出正确词义,在每一轮的注意力计算时都加入词义向量。通过计算词义向量的加权累加和来保存记忆状态 $\boldsymbol{u}^k$,可以表示为

$$\boldsymbol{u}^k = \sum_i \boldsymbol{a}_i^k \boldsymbol{g\_s}_i。\tag{2}$$

根据上一轮的记忆向量 $\boldsymbol{m}^{k-1}$、上下文向量 $\boldsymbol{c}$ 以及记忆状态 $\boldsymbol{u}^k$,采用文献[10]和文献[13]中效果最好的方法更新记忆向量 $\boldsymbol{m}^k$,可以表示为

$$\boldsymbol{m}^k = \mathrm{Relu}(\boldsymbol{W}[\boldsymbol{m}^{k-1}:\boldsymbol{u}^k:\boldsymbol{c}] + \boldsymbol{b}),\tag{3}$$

其中“:”为拼接操作。

1.4　打分模块

打分模块根据记忆模块和上下文表示模块的输出,计算目标词各个词义的分布概率。目标词 w 的第 i 个词义的分数由记忆模块最后一轮的注意力确定,可以表示为

$$\boldsymbol{score}_{g\_s} = \{\boldsymbol{e}_i^k\}。\tag{4}$$

上下文分数由上下文向量经过全连接层得到,可以表示为

$$\boldsymbol{score}_c = \boldsymbol{W}_w\boldsymbol{c} + \boldsymbol{b}_w,\tag{5}$$

式中:$\boldsymbol{W}_w$ 和 $\boldsymbol{b}_w$ 是全连接层的权重矩阵和偏置向量。对于每个目标词 w_t,都有其对应的权重和偏置。

为了综合考虑 $\boldsymbol{score}_{g\_s}$ 和 $\boldsymbol{score}_c$ 对词义分布的影响,引入参数 $\lambda_{w_t} \in [0,1]$。w_t 的词义分布 $\hat{\boldsymbol{y}}$ 可以表示为

$$\hat{\boldsymbol{y}}(w_t) = \mathrm{Softmax}(\lambda_{w_t}\boldsymbol{score}_c + (1-\lambda_{w_t})\boldsymbol{score}_{g\_s})。\tag{6}$$

2　实验数据和方法

2.1　数据集和词典

实验使用的数据集是 SemEval-2007 中英文词义消歧数据集[18],该数据集包含 2 686 条训练语料和 935 条测试语料,40 个用于词义消歧的目标词中包括 21 个动词和 19 个名词,平均每个目标词有 3 个词义。SemEval-2007 中文消歧语料的词义来自《汉语语义词典》(CSD)[19]。CSD 是北京大学构建的语义词典,其中的“释义”字段为该词语的解释,“备注”字段为用法示例,“word”字段为对应的英文单词或短语。SemEval-2007 中词义描述为英文,并且和 CSD 中的“word”字段对应。本文以“word”字段为词义对齐标记,将 CSD 中的“释义”和“备注”字段分别作为词义的解释和例句。对于 CSD 中缺失的释义和例句,根据《现代汉语词典》(DCC)[20]进行补充和完善。补充的释义和例句,使用中科院分词系统 NLPIR(https://github.com/NLPIR-team/NLPIR)进行分词。

2.2　实验方法

为了验证本文模型的效果,以双向长短时记忆网络(Bi-LSTM)为基线模型,对本文模型进行了消融实验,评价指标使用微平均准确率和宏平均准确率[18]。实验中选用文献[21]训练的 300 维词向量,在模型训练中随着模型迭代更新词向量。

本文模型以词语作为基本单位,在上下文表示模块中,以目标词为中心,前后窗口分别设置为 30 个单位;在词义表示模块中,释义+例句的词语长度设置为 40 个单位。长短时记忆网络设置为 1 层,包括 300 个隐藏单元,损失函数为交叉熵。

学习参数设置如下:Batch size 为 100,droupout 为 0.5,迭代次数为 100,学习率为 0.001,学习方法为 Momentum。

3 实验结果与分析

3.1 实验结果

Bi-LSTM 模型将目标词的上下文作为输入,利用 Bi-LSTM+释义、Bi-LSTM+例句、Bi-LSTM+释义+例句方法分别表示 Bi-LSTM 模型中融合目标词释义、例句以及释义+例句信息。不同方法的实验结果如表 1 所示。从表 1 可以看出,本文提出的方法准确率最高。融合语言知识的神经网络模型比仅使用上下文信息的神经网络方法在微平均准确率和宏平均准确率方面均有超过 1%的提高。单独使用例句比单独使用释义在两种准确率上均有提高,说明例句的作用比释义更大。

对本文方法和基线方法的消歧结果进行成对样本 t 检验,P 值为 0.013,说明本文方法与基线方法的消歧结果存在显著差异。

表 1 不同方法的实验结果

Table 1 Results of different methods

方法	微平均准确率/%	宏平均准确率/%
Bi-LSTM(基线方法)	79.89	82.97
Bi-LSTM+释义	81.28	84.34
Bi-LSTM+例句	81.50	84.85
Bi-LSTM+释义+例句(本文方法)	81.82	85.28

Bi-LSTM 方法与本文方法对每个目标词的消歧准确率对比结果表明,本文提出的消歧模型提高了 40%目标词(16/40)的准确率,40%目标词(16/40)的准确率没有变化,20%目标词(8/40)的准确率有所下降。可见,本文方法对大多数目标词的消歧结果有正面的影响。

3.2 释义和例句的作用分析

表 2 列举了词义消歧准确率提升或下降幅度较大的目标词。从表 2 可以看出,Bi-LSTM 模型融合释义和例句对名词、动词的词义消歧准确率均有影响。

本文的模型更容易识别出用释义和例句表示的词义与目标词上下文的语义相似度,从而提高了模型的准确率。例如,目标词“叫”在语料中的词义分别为“ask”“name”“call”和“cry”,其对应的释义分别为“使;让,命令”、“称为;是”、“招呼,呼唤;雇”和“人或动物的发音器官发出较大的声音”,对应的例句分别为“~他早点回家/~人操心”、“~他老李/他没~过你/这~聪明/这~莽撞不~勇敢”、“有人~你/~他去睡午觉/你~老何/车子~了”和“~下去/小鸟~着/小鸡会~了/~坏了嗓子可不好”。在例句“去了三天,蚊香厂却停机三天,叫厂里开机一试,却说机器坏了,所以无法检验”中,Bi-LSTM 模型将“叫”的词义错误地识别为“name”,而本文模型则正确地识别出其词义为“ask”,这是由于本文方法识别出词义“ask”的释义和例句与该句中目标词的上下文有更高的语义相似度。

表 2 消歧准确率提升或下降幅度较大的目标词

Table 2 Target words with higher disambiguation accuracy increase or decrease

目标词	词性	消歧准确率/%		目标词	词性	消歧准确率/%	
		Bi-LSTM	Bi-LSTM+释义+例句			Bi-LSTM	Bi-LSTM+释义+例句
叫	动词	64.10	84.62	出	动词	68.83	61.04
长城	名词	66.67	85.71	面	名词	86.96	82.61
本	名词	88.00	96.00	动摇	动词	93.75	87.50
发	动词	86.11	94.44	说明	动词	94.44	88.89
吃	动词	86.96	95.65	队伍	名词	81.82	72.73
道	名词	66.67	88.89				

外部信息的加入也降低了一些动词和名词的消歧准确率,造成这种情况的主要原因是例句和释义的不完善降低了模型理解词义的能力。例如目标词“出”共有 8 个词义,其中 4 个词义缺少例句。目标词“动摇”

的第 2 个词义的释义用其自身解释为“使动摇”,语义信息不明显。这种例句和释义的不完善使模型不能很好地发现释义和例句与目标词上下文的关系。

3.3　注意力计算轮次的影响

对比了记忆模块中注意力计算轮次对消歧准确率的影响,结果如表 3 所示。从表 3 可以看出,在 3 种语言知识添加的方法中,随着注意力计算轮次的增加,准确率大都有所提升。这是因为随着注意力计算轮次的增加,模型提高了正确词义的注意力。当更新轮次达到 3 次时,3 种语言知识添加的方法大都取得了最高的准确率;随后消歧准确率有所下降,说明多轮注意力虽然能更好地反映目标词上下文与其用释义和例句表达的词义之间的语义关系,但是计算轮次并不是越高越好,需要通过实验确定。

表 3　注意力计算轮次对消歧准确率的影响

Table 3　The effect of attention calculation rounds on disambiguation accuracy

注意力计算轮次	消歧准确率/%		
	Bi-LSTM+释义	Bi-LSTM+例句	Bi-LSTM+释义+例句
第 1 轮	83.09	83.77	84.63
第 2 轮	83.40	84.85	83.87
第 3 轮	84.34	84.31	85.28
第 4 轮	83.74	84.51	83.64

4　小结

本文在神经网络中文词义消歧模型中融合了释义和例句信息,实验结果表明,相对于仅利用上下文信息的神经网络方法,本文模型的宏平均准确率和微平均准确率均提高了约 2%,说明在知识指导下的神经网络模型在词义消歧任务中有明显的作用。下一阶段的工作主要包括以下 3 个方面:第一,利用搜索引擎和已标注的词义语料库[22]扩充例句来提高模型的准确率。第二,改善知识融合方法。本文只是将目标词释义和例句进行简单的拼接,后续的工作可以尝试将释义和例句进行多种方式的结合,把更多的外部知识以及上下文的词性、句法等特征加入到神经网络词义消歧中。第三,完善语言资源的建设。虽然融入语言知识提高了词义消歧的准确率,但是如何解决未登录词以及语言知识不完备的问题还需要进一步的研究。

参考文献:

[1]　宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社, 2013.
ZONG C Q. Statistical natural language processing[M]. Beijing: Tsinghua University Press,2013.

[2]　ZHONG Z, NG H T. It makes sense: a wide-coverage word sense disambiguation system for free text[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics System Demonstrations. Uppsala, 2010: 78-83.

[3]　何径舟,王厚峰. 基于特征选择和最大熵模型的汉语词义消歧[J]. 软件学报, 2010, 21(6):1287-1295.
HE J Z, WANG H F. Chinese word sense disambiguation based on maximum entropy model with feature selection[J]. Journal of software, 2010, 21(6):1287-1295.

[4]　PEDERSEN T. A simple approach to building ensembles of naive Bayesian classifiers for word sense disambiguation[C]// Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference. Hong Kong, 2000: 63-69.

[5]　刘道华,张礼涛,曾召霞,等. 基于正交最小二乘法的径向基神经网络模型[J]. 信阳师范学院学报(自然科学版),2013, 26(3):428-431.
LIU D H, ZHANG L T, ZENG Z X,et al. Radial basis function neural network model based on orthogonal least squares[J]. Journal of Xinyang normal university (natural science edition), 2013,26(3):428-431.

[6]　KÅGEBÄCK M, SALOMONSSON H. Word sense disambiguation using a bidirectional LSTM[EB/OL]. [2019-09-04]. http://arxiv.org/abs/1606.03568.

[7]　RAGANATO A, BOVI C D, NAVIGLI R. Neural sequence learning models for word sense disambiguation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017: 1156-1167.

[8] YANG B S, MITCHELL T. Leveraging knowledge bases in LSTMs for improving machine reading[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver,2017:1436-1446.

[9] LUO F L, LIU T, XIA Q L, et al. Incorporating glosses into neural word sense disambiguation[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018:2473-2482.

[10] SUKHBAATAR S, WESTON J, FERGUS R. End-to-end memory networks[C]// Advances in Neural Information Processing Systems 28. Montreal,2015: 2440-2448.

[11] KUMAR A, IRSOY O, ONDRUSKA P, et al. Ask me anything: dynamic memory networks for natural language processing [C]//Proceedings of the 33rd International Conference on Machine Learning. New York, 2016: 1378-1387.

[12] XIONG C, MERITY S, SOCHER R. Dynamic memory networks for visual and textual question answering[C]//Proceedings of the 33rd International Conference on Machine Learning. New York, 2016: 2397-2406.

[13] CHEN T, XU R F, HE Y L, et al. Improving distributed representation of word sense via WordNet gloss composition and context clustering[J]. Atmospheric measurement techniques, 2015, 4(3):5211-5251.

[14] SASCHA R, HINRICH S. AutoExtend: extending word embeddings to embeddings for synsets and lexemes[EB/OL]. [2019-09-04]. http://arxiv.org/abs/1507.01127.

[15] MILLER G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11):39-41.

[16] POPOV A. Word sense disambiguation with recurrent neural networks[C]// Proceedings of the Student Research Workshop Associated with the International Conference RANLP. Varna, 2017:25-34.

[17] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.

[18] JIN P, WU Y F, YU S W. SemEval-2007 task 05: multilingual Chinese-English lexical sample [C]//Proceedings of the 4th International Workshop on Semantic Evaluations. Prague, 2007: 19-23.

[19] WU Y F, JIN P, ZHANG Y S, et al. A Chinese corpus with word sense annotation[C]//International Conference on Computer Processing of Oriental Languages. Berlin, 2006: 414-421.

[20] 中国社会科学院语言研究所. 现代汉语词典[M]. 北京:商务印书馆, 1983.
INSTITUTE OF LINGUISTICS OF CHINESE ACADEMY OF SOCIAL SCIENCES. Dictionary of contemporary Chinese[M]. Beijing: the Commercial Press,1983.

[21] LI S, ZHAO Z, HU R F, et al. Analogical reasoning on Chinese morphological and semantic relations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, 2018:138-143.

[22] ZAN H Y,CHEN J Y,CHENG X Y,et al. Construction of word sense tagging corpus[C]//Proceedings of the 19th Chinese Lexical Semantics Workshop. Taipei,2018:679-690.

Leveraging Linguistic Knowledge in Neural Network Chinese Word Sense Disambiguation Model

MU Lingling, CHENG Xiaoyu, ZAN Hongying, HAN Yingjie
(*School of Information Engineering*, *Zhengzhou University*, *Zhengzhou* 450001, *China*)

Abstract: A neural network Chinese word sense disambiguation model was proposed. The model leveraged language knowledge to disambiguate senses by using the glosses and samples of target words in bidirectional long-term and short-term memory network. Experimental results of this model on the SemEval-2007 dataset of Chinese-English lexicon sample showed that the macro and micro average-accuracy of the model were increased by 2.31% and 1.93% respectively compared to the baseline model after leveraging language knowledge. The results demonstrated that leveraging language knowledge with neural network models was helpful for improving Chinese word sense disambiguation.

Key words: word sense disambiguation; dictionary; neural network; language resource

(责任编辑:孔 薇)