doi:10.11887/j.cn.201803023 http://journal.nudt.edu.cn

关系抽取中远监督错误标注消除*

汝承森,唐晋韬,谢松县,李莎莎,王 挺
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:目前远监督方法被广泛应用于关系抽取任务。然而,远监督方法中存在大量错误标注现象,给远监督方法的学习效果带来了很大的影响。提出利用语义Jaccard度量关系短语与依存词间语义相似性的错误标注消除方法。消除错误标注后的训练数据用于训练模型,完成关系抽取。实验结果表明:该方法可以有效消除错误标注,提高关系抽取的性能。

关键词:关系抽取;远监督;错误标注;语义相似性
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2018)03-148-05

Reducing wrong labels in distant supervision for relation extraction

RU Chengsen, TANG Jintao, XIE Songxian, LI Shasha, WANG Ting
(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: Distant supervision has been widely used for relation extraction recently. In the distant supervision, many labels may to wrongly marked, which exerts a bad impact on relation extraction. A method to reduce wrong labels was introduced by using the semantic Jaccard to measure semantic similarity between the relation phrases and the dependency terms. The training data after reducing wrong labels was used to train the relation extractors. The experimental results show that the proposed method can effectively reduce wrong labels and improve the relation extraction performance compared with the state-of-art methods.

Key words: relation extraction; distant supervision; wrong labels; semantic similarity

当今时代,信息呈现爆炸式增长,能够快速准确地从海量信息中获取用户所需要的信息显得尤为重要。信息抽取技术[1-2]的出现为用户解决了这一难题。关系抽取是信息抽取的关键技术之一,是一个从文本中抽取结构化信息的过程[3],对于问答系统、机器阅读以及知识图谱等应用具有重要意义。但是,关系抽取方法通常面临缺少标注数据问题[4]。标注数据需要耗费大量人力物力。为缓解标注语料不足问题,Mintz等[5]利用远监督方法进行关系抽取。如果一个句子包含的实体对与知识库中已有关系实例的实体对相同,远监督方法将该句子标注为对应关系的实例,具体流程如图1所示。

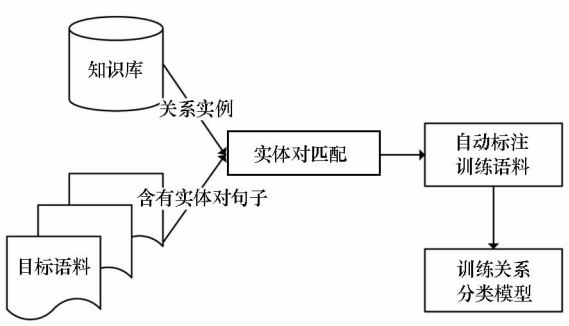

图1 远监督关系抽取流程

Fig. 1 Process of distant supervision relation extraction

远监督方法可以自动标注训练语料,节省了大量人力物力。但是,由于两个实体间的关系可能不止一种,这样就会导致错误标注现象。为消除错误标注,Riedel[6]与Hoffmann[7]等假设同一实体对的所有出现中至少有一个是特定关系的正确描述,使用多实例学习消除错误标注训练正例的影响。他们的工作没有直接消除错误标注训练正例。当实体对只在语料中出现了一次并且没有描述目标关系时,上述方法就会失效。为了直接消除错误标注训练正例,Takamatsu等[8]提出了一种生成模型方法,依据标注语料评估关系对应模式的概率,消除错误标注。Han等[9]提出了一

* **收稿日期**:2016-11-25
基金项目:国家自然科学基金资助项目(61472436,61532001,61303190)
作者简介:汝承森(1988—),男,山东聊城人,博士研究生,E-mail:ruchengsen@nudt.edu.cn;
王挺(通信作者),男,教授,博士,博士生导师,E-mail:tingwang@nudt.edu.cn

种基于局部子空间的方法,利用语义一致性区分正确与错误标注。他们的方法依赖于大规模的标注数据,当标注数据不足或标注数据中出现大量错误标注时效果就会受到影响。

为避免错误标注消除效果受到标注数据规模、质量影响,本文提出了一种基于语义相似性的方法,以提高远监督学习的效果。知识库利用关系短语描述各种关系类型,而实体间关系由依存路径上的词语(依存词)描述。因此,可以通过度量关系短语与依存词间的语义相似度判断句子是否为正确标注。语义相似度越高,正确标注的概率越大;语义相似度越低,正确标注的概率越小。关系短语与依存词的语义表示是度量相似度需要解决的首要问题。近年来,词向量[10]被广泛用于表示单语语义,不需要进行语义扩充就可以取得很好的效果。基于此,本文利用词向量表示关系短语以及依存词的语义,将关系短语看作一个句子,属于同一句子的所有依存词看作另外一个句子,利用语义 Jaccard[11]度量关系短语与依存词的语义相似性。当与实体对对应的语义 Jaccard 值大于某个相似度阈值时,该实体对所在的句子为正确标注实例,否则,为错误标注实例。

1　维基百科与 YAGO 知识库

从图 1 可知,要利用远监督进行关系抽取需要选择合适的目标语料与知识库。

维基百科是一个基于 WEB2.0 技术的多语言百科全书,已成为互联网上最大的、使用最广泛的开放式电子百科全书,包含了数百万的文档语料,质量上和数量上都有其他语料库无法比拟的优势。作为一个领域覆盖广泛,知识增长和更新速度相当快的免费百科全书,维基百科为抽取语义关系知识、构建知识库等应用提供了丰富的、可靠的、低成本的内容资源。基于此,使用英文版维基百科作为目标语料。

YAGO[12]是一个在线的知识库,可以自由访问,数据主要来源于维基百科。YAGO 从维基百科中自动抽取信息,并使用 WordNet 进行结构化处理,形成了覆盖面较全、数据质量较高的大规模语义知识库。YAGO 规定了关系类型与实体种类。在 YAGO 中,两个实体与一个关系组成的三元组成为一个事实,称为关系实例。

由于 YAGO 中的几乎所有信息源自维基百科,相对于其他知识库,YAGO 中关系实例出现在维基百科句子中的概率更大,更容易获得标注数据。为获取足够多的标注数据,选用 YAGO2s[13]作为知识库。

为研究解决错误标注问题的方法,从 YAGO2s 选择了四种容易被混淆的关系类型进行实验:每种关系在语料中有多种表述并且其包含的实体对可能有多种关系,容易出现错误标注现象。这四种关系分别为 was_born_in、died_in、is_affiliated_to 以及 created。表 1 详细列出了每种关系包含的关系实例数量。

表 1　关系信息介绍

Tab. 1　Introduction of the relation types

关系类型	关系实例规模
was_born_in	218 757
died_in	54 174
is_affiliated_to	497 263
created	278 455

2　错误标注消除

如果实体对间存在关系,位于实体对间依存路径上的词语(依存词)可以作为识别实体间关系的特征,体现关系语义。同时,知识库以关系短语表示特定关系类别。因此,可以通过度量实体对间依存词与关系短语的语义相似度,判断利用远监督方法标注的原始训练正例是否为指定关系的正确标注,以达到消除错误标注的目的。此时,筛选正确正向标注的关键在于度量实体间依存词与关系短语的语义相似度。

2.1　依存词

Wu 等[14]指出,实体间存在一条最短依存路径,并且路径上的词语(依存词)代表实体间的关系。虽然依存词可以表明实体间存在关系,但是它们不一定完全获取了关系语义。为真正获取关系语义,可利用扩展依存路径(通过增加形容词或副词修饰语到最短依存路径得到的树状结构)上的词语。考虑下面句子:David was not born in Bethlehem。

图 2、图 3 分别给出了实体对 David 与 Bethlehem 间的最短依存路径与扩展依存路径。在上述句子中,实体对间不存在关系 was_born_in。单词 born 是从 David 到 Bethlehem 的最短依存路径上唯一的依存词。仅仅依据单词 born 不足以判断句子没有描述关系 was_born_in。为了获取关系的真正语义,应该考虑单词 born 在扩展依存路径上的修饰词 was、not 和 in。为此引入修

饰语和扩展依存词的概念,以准确表达实体之间的语义关联信息。

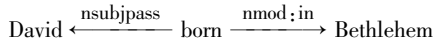

图2 最短依存路径

Fig. 2 Shortest dependency path

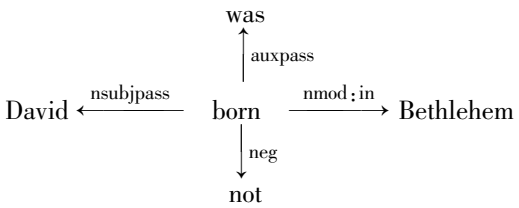

图3 扩展依存路径

Fig. 3 Expanded dependency path

定义1 修饰语:在扩展依存路径上,用于修饰依存词的形容词或副词,例如,was、not 和 in 都是依存词 born 的修饰语。

定义2 扩展依存词:依存词以及它的所有修饰语构成的整体,例如,词链 was not born in 构成一个扩展依存词。

2.2 利用词向量进行词汇语义表示

近年来,词向量被广泛用于表示单词语义,不需要进行语义扩充就可以取得很好的效果。基于此,利用词向量表示关系短语以及依存词的语义。在实际应用中,通常以向量的形式使用词向量,向量中的每一维对应单词的一个语义或者语法特征。在词向量中,单词被映射到多维空间中,两个词义越相近的单词在空间中的位置距离越近。现在有很多训练好的词向量,可以在研究中直接使用。本文使用 Turian 等[15]训练的词向量。

一个扩展依存词包含一个依存词以及该依存词的所有修饰语。扩展依存词的语义表示可以通过式(1)计算获得。

$$\boldsymbol{U}_i = \alpha \boldsymbol{V}_i + \beta \sum_{j=1}^{m} \boldsymbol{M}_{ij} \tag{1}$$

式中,$\boldsymbol{U}_i \in \mathbf{R}^n$ 表示第 i 个扩展依存词,$\boldsymbol{V}_i \in \mathbf{R}^n$ 表示第 i 个扩展依存词中包含的依存词,$\boldsymbol{M}_{ij} \in \boldsymbol{R}^n$ 表示第 i 个依存词的第 j 个修饰语,n 表示词向量的维度,α、β 分别表示依存词与修饰语的权重,m 表示修饰语的数量。$\boldsymbol{V}_i$ 与 $\boldsymbol{M}_{ij}$ 对应的值可以通过查询词向量词典直接获得。为了区分依存词与其修饰语的不同影响,将 α 的权重设置为2,将 β 的权重设置为1。

对于每种关系,关系短语只包含一个核心词 $\boldsymbol{V}_{\mathrm{ck}}$ 以及若干个修饰语 $\boldsymbol{M}_{\mathrm{kj}}$,可以利用式(2)计算关系短语对应的向量 $\boldsymbol{RP}_{\mathrm{k}}$。在式(2)中,核心词与修饰语的权重与式(1)相同。

$$\boldsymbol{RP}_{\mathrm{k}} = \alpha \boldsymbol{V}_{\mathrm{ck}} + \beta \sum_{j=1}^{l} \boldsymbol{M}_{\mathrm{kj}} \tag{2}$$

式中,l 表示关系短语中所含修饰语的数量。

2.3 利用语义 Jaccard 进行相似性度量

Jaccard[16]通常被用于度量句子间的语义相似度。原始的 Jaccard 只使用文本匹配的方式进行相似度度量,单一的文本匹配不能提供足够的信息,限制了它的使用。为解决这个问题,Zhang[7]用词向量表示句子中单词的语义,提出了语义 Jaccard。语义 Jaccard 的定义可表示为:

$$SemJac(X,Y) = \frac{s}{s+d} \tag{3}$$

$$s = \sum_{m} \cos(x_{\mathrm{sim}}, y_{\mathrm{sim}}), \cos(x_{\mathrm{sim}}, y_{\mathrm{sim}}) \geqslant \delta \tag{4}$$

$$d = \sum_{n} [1 - \cos(x_{\mathrm{dif}}, y_{\mathrm{dif}})], \cos(x_{\mathrm{dif}}, y_{\mathrm{dif}}) < \delta \tag{5}$$

$$m + n = \min(|X|, |Y|) \tag{6}$$

其中,x_{sim}、x_{dif} 表示属于同一句子的不同 n-grams,y_{sim}、y_{dif} 表示属于另一句子的不同 n-grams。在式(3)中,分子表示语义相同部分,分母的第一部分与分子相同,第二部分表示语义不同部分。δ 表示相似度阈值,取值范围为 -1 ~ 1。

可以将关系短语看作一个句子,将属于同一句子的所有扩展依存词看作另外一个句子。一个关系短语作为一个 n-gram,一个扩展依存词作为一个 n-gram。由于一个关系短语对应的 n-gram 数量始终为一,因此可以用式(7)代替式(3) ~ (6)。

$$SemJac(X,Y) = \begin{cases} 1 & \exists \cos(x_i, y_j) \geqslant \delta \\ 0 & \forall \cos(x_i, y_j) < \delta \end{cases} \tag{7}$$

将消除错误标注问题变为判断是否存在一个扩展依存词,使得该扩展依存词与关系短语的语义相似度不小于相似度阈值 δ。

3 实验

使用英文版维基百科作为目标语料,选用 YAGO2s 作为知识库。实验中,使用了 850 000 篇英文维基百科文章:800 000 篇用于训练,50 000 篇用于测试,每篇文章所含单词数不小于 5000。使用命名实体工具 Stanford Named Entity Recognizer 对句子进行命名实体识别,使用 Stanford Parser 对句子进行句法分析。

从 YAGO2s 选择了四种容易被混淆的关系类型进行实验:每种关系在语料中有多种表述并且

其包含实体对可能有多种关系,容易出现错误标注现象。这四种关系分别为 was_born_in、died_in、is_affiliated_to 以及 created。在实验中,根据四种关系在 YAGO2s 中的关系实例进行数据标注,得到的原始训练正例详情见表 2。表 2 中的错误率是通过对每种关系中的原始训练正例进行随机采样,每次取 100 个样例,人工判断样例是否正确,计算每次采样的错误率,随机采样 10 次,对错误率求平均值得到的。四种关系平均错误率达到 74.1%。

表 2　原始训练正例情况

Tab. 2　Details of the original positive examples

关系类型	原始训练正例数量	错误率/%
was_born_in	37 017	81.7
died_in	22 405	95.6
is_affiliated_to	29 817	50.2
created	25 633	68.8

对于用于测试的维基百科文章,只保留包含实体数量大于 2 的句子用于测试。最终,测试集含有 11 000 000 个句子。关系抽取效果受到训练样例数量、质量的影响。虽然,相似度阈值越高,消除错误标注后的训练样例质量越高,但是训练样例的数量越少。设置相似度阈值时,需要兼顾训练样本的数量与质量。因此,测试集被分成两部分:一部分作为验证集,含有 400 000 个句子,用于选择最佳相似度阈值,以确保取得最好的关系抽取效果;另一部分作为测试集,含有 700 000 个句子,用于评价关系抽取性能。验证集上的关系抽取效果显示,关系 was_born_in、died_in、is_affiliated_to 以及 created 的最佳相似度阈值分别为 0.7、0.7、0.4 与 0.4。

在关系抽取任务上,基于远监督方法的关系抽取有 held-out 评价和人工评价两种评价方式。held-out 评价需要将知识库中每种关系的所有关系实例分成两部分:一部分用于自动标注训练样例;剩余部分测试新发现的关系实例。人工评价至少需要三个人同时评判每个新发现关系实例是否正确,按照投票原则确定最终评判结果。由于 held-out 评价方式经常面临知识库不完备问题,不能全面反映关系抽取效果,采用人工评价方式对每种关系中置信度最高的 N 个(top$N$)新发现关系实例进行评判。

实验中所有模型使用 Mintz 等[5]开发的句法特征以及词法特征。对表 3 中的方法进行了关系抽取效果对比。在所有方法中,训练反例与训练正例规模相同。

表 3　实验方法介绍

Tab. 3　Introduction of methods

方法名称	使用模型	训练数据
FLR	逻辑回归	使用最佳相似度阈值消除错误标注后的训练数据
FMultiR	多示例学习	使用最佳相似度阈值消除错误标注后的训练数据
LR	逻辑回归	原始标注数据
MultiR	多示例学习	原始标注数据

方法 LR 与 FLR 使用逻辑回归模型,并利用 L-BFGS 方法对模型优化。MultiR 与 FMultiR 使用多示例学习模型,实验设置与 Hoffmann 等[9]的一致。FLR、FmultiR 采用的训练数据为消除错误标注后的数据,而 LR、MultiR 采用了原始标注数据。分别比较方法 FLR、FMultiR、LR 以及 MultiR 在所有关系上 top50、top100 以及 top200 的准确率,见表 4。

表 4　各实验方法效果

Tab. 4　Performance of methods

关系	方法	top50	top100	top200
was_born_in	FLR	0.68	0.8	0.845
	FMultiR	0.88	0.89	0.845
	LR	0.44	0.29	0.21
	MultiR	0.18	0.38	0.435
died_in	FLR	0.8	0.49	0.35
	FMultiR	0.9	0.76	0.58
	LR	0	0.01	0.005
	MultiR	0.18	0.24	0.265
is_affiliated_to	FLR	0.82	0.91	0.955
	FMultiR	0.64	0.8	0.895
	LR	0.24	0.18	0.09
	MultiR	0.18	0.27	0.36
created	FLR	0.12	0.15	0.125
	FMultiR	0.18	0.21	0.22
	LR	0.2	0.14	0.095
	MultiR	0.04	0.09	0.17

如表 4 所示,方法 FMultiR 与 FLR 性能明显优于其他方法。通过对比可以发现,方法 FLR 比 LR 在四种关系上的 top50、top100、top200 平均准确率分别提高了 175%、279%、411%,而方法

FMultiR 比 MultiR 在四种关系上的 top50、top100、top200 平均准确率分别提高了 348%、171%、107%。这说明利用语义相似度进行错误标注消除的方法是有效的,可以提高关系抽取的效果。

方法 FMultiR 整体性能最好,方法 MultiR 的整体性能特别是在 top100 与 top200 上明显优于方法 LR。这表明,使用多实例学习可以在一定程度上减少错误标注的影响。

方法 MultiR 性能低于 FLR,这表明,虽然多实例学习可以在一定程度上降低错误标注的影响,但是当错误标注比较频繁时,还是会受到错误标注的影响。

LR 的性能是最差的。从表 4 中可以看出,LR 方法中几乎不能抽取关系 died_in。通过随机抽样检查关系 died_in 的原始训练正例,发现关系 died_in 超过 95% 的原始训练正例是错误标注的,在训练关系抽取模型时会引入大量噪音,这严重影响了关系抽取的性能。

四种关系类型中,使用原始标注数据训练模型时,关系 died_in 的效果是最差的。通过比较发现,关系 died_in 的原始训练正例的错误率是最高的。这进一步表明,训练数据的质量直接关系到关系抽取的性能。同时,进行错误标注消除,使用过滤数据训练模型后,关系 died_in 的效果提升也是最明显的。这说明利用语义相似度进行错误标注消除的方法是有效的,确实提高了关系抽取的效果。

4 结论

本文提出一种基于语义相似度的远监督错误标注消除方法。本方法利用语义 Jaccard 结合词向量度量关系短语与句子中实体对间依存词的语义相似度。通过设置相似度阈值,利用语义相似度消除错误标注。实验结果表明本方法有效地消除了错误标注,显著提高了关系抽取的效果。

参考文献(References)

[1] 于龙, 尹浩. 站点主题结构与导航归纳技术[J]. 国防科技大学学报, 2012, 34(5): 90 - 95.
YU Long, YIN Hao. Website topic structure and navigation induction[J]. Journal of National University of Defense Technology, 2012, 34(5): 90 - 95. (in Chinese)

[2] 钟志农, 刘方驰, 吴烨, 等. 主动学习与自学习的中文命名实体识别[J]. 国防科技大学学报, 2014, 36(4): 82 - 88.
ZHONG Zhinong, LIU Fangchi, WU Ye, et al. Chinese named entity recognition combined active learning with self-training[J]. Journal of National University of Defense Technology, 2014, 36(4): 82 - 88. (in Chinese)

[3] Banko M, Cafarella M J, Soderl S, et al. Open information extraction from the web[C]//Proceedings of the International Joint Conferences on Artificial Intelligence, 2007: 2670 - 2676.

[4] 杨博, 蔡东风, 杨华. 开放式信息抽取研究进展[J]. 中文信息学报, 2014, 28(4): 1 - 11.
YANG Bo, CAI Dongfeng, YANG Hua. Progress in open information extraction[J]. Journal of Chinese Information Processing, 2014, 28(4): 1 - 11. (in Chinese)

[5] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, 2009: 1003 - 1011.

[6] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[C]//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, 2010: 148 - 163.

[7] Hoffmann R, Zhang C L, Ling X, et al. Knowledge-based weak supervision for information extraction of overlapping relations[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011: 541 - 550.

[8] Takamatsu S, Sato I, Nakagawa H. Reducing wrong labels in distant supervision for relation extraction[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012: 721 - 729.

[9] Han X P, Sun L. Semantic consistency: a local subspace based method for distant supervised relation extraction[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 718 - 724.

[10] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3: 1137 - 1155.

[11] Zhang J L. One of the poor semantic processing toolbox: the semantic Jaccard[EB/OL]. [2016 - 08 - 12]. http://blog.csdn.net/ malefactor/article/details/50471118.

[12] Suchanek F M, Kasneci G, Weikum G. YAGO: a core of semantic knowledge unifying wordnet and Wikipedia[C]//Proceedings of the 16th International Conference on World Wide Web, 2007: 697 - 706.

[13] Biega J, Kuzey E, Suchanek F M. Inside YAGO2s: a transparent information extraction architecture[C]//Proceedings of the 22nd International Conference on World Wide Web, 2013: 325 - 328.

[14] Wu F, Weld D S. Open information extraction using Wikipedia[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 118 - 127.

[15] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 384 - 394.

[16] Real R, Vargas J M. The probabilistic basis of jaccard's index of similarity[J]. Systematic Biology, 1996, 45(3): 380 - 385.