基于深度学习的关系抽取

作者：林衍凯、刘知远（清华大学）

【引言】

信息抽取旨在从大规模非结构或半结构的自然语言文本中抽取结构化信息。关系抽取是其中的重要子任务之一，主要目的是从文本中识别实体并抽取实体之间的语义关系。例如，句子“Bill Gates is the founder of Microsoft Inc.”中包含一个实体对(Bill Gates, Microsoft Inc.)，这两个实体对之间的关系为 Founder。

现有主流的关系抽取技术分为有监督的学习方法、半监督的学习方法和无监督的学习方法三种：

1、有监督的学习方法将关系抽取任务当做分类问题，根据训练数据设计有效的特征，从而学习各种分类模型，然后使用训练好的分类器预测关系。该方法的问题在于需要大量的人工标注训练语料，而语料标注工作通常非常耗时耗力。

2、半监督的学习方法主要采用 Bootstrapping 进行关系抽取。对于要抽取的关系，该方法首先手工设定若干种子实例，然后迭代地从数据从抽取关系对应的关系模板和更多的实例。

3、无监督的学习方法假设拥有相同语义关系的实体对拥有相似的上下文信息。因此可以利用每个实体对对应上下文信息来代表该实体对的语义关系，并对所有实体对的语义关系进行聚类。

与其他两种方法相比，有监督的学习方法能够抽取更有效的特征，其准确率和召回率都更高。因此有监督的学习方法受到了越来越多学者的关注，本文也将重点介绍该类方法。

深度学习是机器学习研究的热点之一领域，其主要思想是模拟人脑神经网络建立学习模型，从语音、图像或文本等不同数据中学习有用信息。典型的深度学习方法包括卷积神经网络（convolutional neural networks，CNN）和循环神经网络（recurrent neural networks，RNN），这些模型在文本分类、机器翻译、智能问答等方向都取得了显著的效果。那么，近年来深度学习技术在关系抽取领域的进展如何，关系抽取技术未来的研究趋势是什么？本文将就这些问题加以详细的阐述。

【基于有监督学习的关系抽取】

有监督的关系抽取系统通常需要大量人工标注的训练数据，从训练数据中自动学习关系对应的抽取模式。有监督关系抽取方法主要包括：基于核函数的方法[Zhao and Grishman 2005; Bunescu and Mooney 2006]，基于逻辑回归的方法[Kambhatla 2004]，基于句法解析增强的方法[Miller et al. 2000]和基于条件随机场的方法[Culotta et al. 2006]。然而，阻碍这些系统效果继续提升的主要问题在于，人工标注训练数据需要花费大量的时间和精力。

针对这个局限性，Mintz 等人[Mintz et al. 2009]提出了远程监督（Distant Supervision）的思想。作者们将纽约时报新闻文本与大规模知识图谱 Freebase（包含 7300 多个关系和超过 9 亿的实体）进行实体对齐。远程监督假设，一个同时包含两个实体的句子蕴含了该实体对在 Freebase 中的关系，并将该句子作为该实体对所对应关系的训练正例。作者在远程监督标注的数据上提取文本特征并训练关系分类模型，有效解决了关系抽取的标注数据规模问题。之后许多研究者从各个角度对远程监督技术提出了改进方案。例如 Takamatsu 等人[Takamatsu et al. 2012]改进了实体对齐的技术，降低了数据噪音，提高了关系抽取的总体效果。Yao 等人[Yao et al. 2010]提出了基于无向图模型的关系抽取方法。Riedel 等人[Riedel et al. 2010]则增强了远程监督的假设，与 [Mintz et al.2009]相比错误率减少了 31%。

以上远程监督技术都假设一个实体对只对应一种关系。但是，很多实体之间具有多种关系。例如，“Steve Jobs founded Apple”和“Steve Jobs is the CEO of Apple”。因此，Hoffmann 等人[Hoffmann et al. 2011]提出采用多实例多标签（Multi-Instance Multi-label）方法来对关系抽取进行建模，刻画一个实体对可能存在多种关系的情况。类似地，Surdeanu 等人[Surdeanu et al. 2012]也提出利用多实例多标签和贝叶斯网络来进行关系抽取。

【基于深度学习的关系抽取】

现有的有监督学习关系抽取方法已经取得了较好的效果，但它们严重依赖词性标注、句法解析等自然语言处理标注提供分类特征。而自然语言处理标注工具

往往存在大量错误，这些错误将会在关系抽取系统中不断传播放大，最终影响关系抽取的效果。

最近，很多研究人员开始将深度学习的技术应用到关系抽取中。[Socher et al. 2012] 提出使用递归神经网络来解决关系抽取问题。该方法首先对句子进行句法解析，然后为句法树上的每个节点学习向量表示。通过递归神经网络，可以从句法树最低端的词向量开始，按照句子的句法结构迭代合并，最终得到该句子的向量表示，并用于关系分类。该方法能够有效地考虑句子的句法结构信息，但同时该方法无法很好地考虑两个实体在句子中的位置和语义信息。

[Zeng et al. 2014] 提出采用卷积神经网络进行关系抽取。他们采用词汇向量和词的位置向量作为卷积神经网络的输入，通过卷积层、池化层和非线性层得到句子表示。通过考虑实体的位置向量和其他相关的词汇特征，句子中的实体信息能够被较好地考虑到关系抽取中。后来，[Santos et al. 2015]还提出了一种新的卷积神经网络进行关系抽取，其中采用了新的损失函数，能够有效地提高不同关系类别之间的区分性。

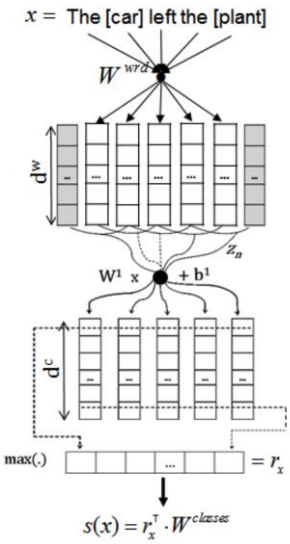

图 1 CNN 进行关系分类示意图[Santos et al. 2015]

[Miwa et al. 2016] 提出了一种基于端到端神经网络的关系抽取模型。该模型使用双向 LSTM（Long-Short Term Memory，长短时记忆模型）和树形 LSTM 同时对实体和句子进行建模。目前，基于卷积神经网络的方法在关系抽取的标准数据集 SemEval-2010 Task 8 上取得了最好的效果。

上面介绍的神经网络模型在人工标注的数据集上取得了巨大成功。然而，与之前基于特征的关系抽取系统类似，神经网络关系抽取模型也面临着人工标注数据较少的问题。对此，[Zeng et al. 2015]尝试将基于卷积神经网络的关系抽取模型扩展到远程监督数据上。[Zeng et al. 2015]假设每个实体对的所有句子中至少存在一个句子反映该实体对的关系，提出了一种新的学习框架：以实体对为单位，对于每个实体对只考虑最能反映其关系的那个句子。该方法在一定程度上解决了神经网络关系抽取模型在远程监督数据上的应用，在 NYT10 数据集上取得了远远高于基于特征的关系抽取模型的预测效果。但是，该方法仍然存在一定的缺陷：该模型对于每个实体对只能选用一个句子进行学习和预测，损失了来自其他大量的有效句子的信息。

我们有没有可能把实体对对应的有噪音的句子过滤掉，然后利用所有有效句子进行学习和预测呢？ [Lin et al. 2016] 提出了一种基于句子级别注意力机制的神经网络模型来解决这个问题，该方法能够根据特定关系为实体对的每个句子分配权重，通过不断学习能够使有效句子获得较高的权重，而有噪音的句子获得较小的权重。与之前的模型相比，该方法效果取得较大提升。我们也将相关代码发布在 Github 上：https://github.com/thunlp/NRE。

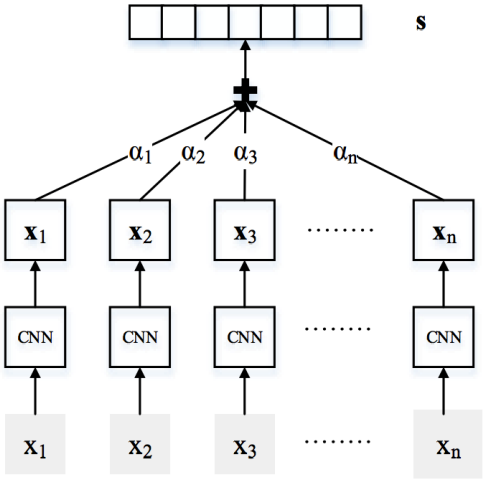

图 2 基于句子级别注意力机制的神经网络模型示意图 [Lin et al. 2016]

【总结及未来趋势】

近年来，深度学习在自然语言处理领域的很多方向取得了巨大成功，本文以关系抽取为例，介绍了如何利用深度学习的语义表示和学习能力，自动地从训练数据中学习分类特征，从而取得比传统方法更好的关系抽取效果。当然，关系抽取系统性能还有很大提升空间，仍然有很多问题亟待解决。

首先，基于句法树的树形 LSTM 神经网络模型在关系抽取上取得了不错的效果，这说明句法信息的引入对于关系抽取有一定帮助。然而，目前的句法分析仍然存在较多错误，在考虑句法信息的同时也引入了大量错误噪音。有研究表明，如果对于一个句子考虑其最可能的多个句法分析树，分析结果准确率可以得到较大提升。因此，一个重要的研究方向是，如何有效地将句子的多个可能句法树信息结合起来，用于关系抽取。

其次，目前的神经网络关系抽取主要用于预先设定好的关系集合。而面向开放领域的关系抽取，仍然是基于模板等比较传统的方法。因此，我们需要探索如何将神经网络引入开放领域的关系抽取，自动发现新的关系及其事实。此外，对现有神经网络模型如何对新增关系和样例进行快速学习也是值得探索的实用问题。

最后，目前关系抽取主要基于单语言文本。事实上，人类知识蕴藏于不同模态和类型的信息源中。我们需要探索如何利用多语言文本、图像和音频信息进行关系抽取。

【参考文献】

1. Zhao, Shubin, and Ralph Grishman. "Extracting relations with integrated information using kernel methods." In Proceedings of ACL, 2005.

2. Mooney, Raymond J., and Razvan C. Bunescu. "Subsequence kernels for relation extraction." In Proceedings of NIPS, 2005.

3. Kambhatla, Nanda. "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations." In Proceedings of ACL, 2004.

4. Miller, Scott, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. "A novel use of statistical parsing to extract information from text." In Proceedings of NAACL, 2000.

5. Culotta, Aron, Andrew McCallum, and Jonathan Betz. "Integrating

probabilistic extraction models and data mining to discover relations and patterns in text." In Proceedings of HLT-NAACL, 2006.

6. Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. "Distant supervision for relation extraction without labeled data." In Proceedings of ACL-IJCNLP, 2009.

7. Takamatsu, Shingo, Issei Sato, and Hiroshi Nakagawa. "Reducing wrong labels in distant supervision for relation extraction." In Proceedings of ACL, 2012.

8. Yao, Limin, Sebastian Riedel, and Andrew McCallum. "Collective cross-document relation extraction without labelled data." In Proceedings of EMNLP, 2010.

9. Riedel, Sebastian, Limin Yao, and Andrew McCallum. "Modeling relations and their mentions without labeled text." In ECML/PKDD, 2010.

10. Hoffmann, Raphael, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. "Knowledge-based weak supervision for information extraction of overlapping relations." In Proceedings of ACL-HLT, 2011.

11. Surdeanu, Mihai, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. "Multi-instance multi-label learning for relation extraction." In Proceedings of EMNLP-CoNLL, 2012.

12. Makoto Miwa, Mohit Bansa. “End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures” In Proceedings of ACL, 2016.

13. Socher, Richard, et al. "Semantic compositionality through recursive matrix-vector spaces." Proceedings of EMNLP-CoNLL, 2012.

14. Santos, Cicero Nogueira dos, Bing Xiang, and Bowen Zhou. "Classifying relations by ranking with convolutional neural networks." In Proceedings of ACL, 2015.