ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-201806.00128

Distant Supervision Relation Extraction Based on GRU and Attention Mechanism Postprint

Authors: Huang Zhaowei, always-on, Bin Chenzhong, Sun Yanpeng, Sun Lei

Date: 2018-06-19T00:00:00+00:00

Abstract

With the advancement of deep learning, an increasing number of deep learning models have been employed for relation extraction tasks. However, traditional deep learning models are unable to address the issue of long-distance dependencies. Simultaneously, distant supervision inevitably introduces noisy labels. To tackle these two challenges, we propose a distant supervision relation extraction method based on GRU (gated recurrent unit) and attention mechanisms. First, GRU neural networks are utilized to extract textual features, thereby resolving the long-distance dependency problem. Next, a sentence-level attention mechanism is constructed over entity pairs to reduce the weights of noisy sentences. Finally, on a real-world dataset, we demonstrate through precision and recall calculations and PR curve plotting that the proposed method achieves significant improvements compared to several existing approaches.

Full Text

Distant Supervision Relationship Extraction Based on GRU and Attention Mechanism

Huang Zhaowei, Chang Liang, Bin Chenzhong†, Sun Yanpeng, Sun Lei (Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China)

Abstract: With the development of deep learning, an increasing number of deep learning models have been applied to relation extraction tasks. However, traditional deep learning models cannot solve long-distance dependency problems, while distant supervision inevitably generates incorrect labels. To address these two issues, this paper proposes a distant supervision relationship extraction method based on GRU (Gated Recurrent Unit) and attention mechanisms. First, a GRU neural network is employed to extract text features, thereby solving long-distance dependency problems. Next, a sentence-level attention mech-

anism is constructed on entity pairs to reduce the weight of noisy sentences. Finally, on a real dataset, by calculating precision and recall and plotting PR curves, the proposed method demonstrates significant improvements compared to existing approaches.

Keywords: deep learning; distant supervision; GRU; attention mechanism

0 Introduction

Although existing knowledge bases already contain substantial knowledge, the advent of the big data era has made this knowledge insufficient to meet human needs. Consequently, there is an urgent demand to extract structured data from unstructured text. Relation extraction [1] is not only a fundamental task in information extraction but also crucial for constructing and completing knowledge graphs. The primary objective of this task is to mine semantic relationships between entities from textual content [2], generating relational data from plain text, which represents a key challenge in natural language processing (NLP). This task can be formally described as: given a text S, determine the relationship category r between two target entity pairs.

In recent years, knowledge bases such as Freebase [3], DBpedia [4], and YAGO [5] have been established and widely applied in NLP tasks including search, recommendation, and question answering. These knowledge bases consist of factual triples, such as (Yao Ming, spouse, Ye Li). Despite the vast knowledge contained in these repositories, the big data era demands even more. This paper proposes a distant supervision relationship extraction method that combines GRU (Gated Recurrent Unit) neural networks with attention mechanisms (GRU_ATT). The approach employs distant supervision to avoid time-consuming and labor-intensive manual dataset construction, utilizes GRU models to overcome the long-distance dependency limitations [6] of traditional deep learning models, and introduces sentence-level attention mechanisms to effectively control the impact of noisy data on experimental results. Finally, GRU_ATT is evaluated on real datasets, and experimental results demonstrate significant improvements in relation extraction compared to existing methods. The main contributions of this paper are: (a) a method for constructing attention mechanisms at the sentence level, and (b) a relationship extraction approach that integrates attention mechanisms with GRU.

1 Related Work

Traditional relation classification methods are primarily based on pattern matching. For instance, Harabagiu et al. [7] proposed a method for relation classification by combining lexical and semantic relationships, while Kambhatla [8] introduced a logistic regression-based approach for relation extraction. These pattern-matching methods have achieved good performance but suffer from several drawbacks. First, many traditional NLP systems are required to extract high-level features such as part-of-speech tags, shortest dependency paths, and

named entities, leading to increased computational costs and additional propagation errors. Second, due to low coverage across different training datasets, these methods exhibit poor generalization.

Most existing supervised relation extraction methods require large amounts of labeled training data, which is extremely time-consuming and labor-intensive. In 2009, Mintz et al. [9] first proposed distant supervision, which automatically generates training datasets by aligning triple knowledge bases (KB) with text. Their assumption states that if two entities have a relationship in the KB, then all sentences containing these two entities express this relationship. For example, the triple fact (Yao Ming, spouse, Ye Li) in the KB would lead distant supervision to treat all sentences containing both entities as positive examples of this relationship. Although distant supervision is an effective strategy for automatically labeling training data, it introduces incorrect labeling issues. For instance, sentences like "Yao Ming was invited to participate in Tencent Sports Celebrity Game, Ye Li expressed great support" and "Yao Ming and Ye Li appeared together at Shanghai Pudong International Airport"contain both entities but do not express the "spouse" relationship, yet are still treated as positive examples in distant supervision.

To address this, Riedel et al. [10] adopted a multi-instance learning approach, treating all sentences labeled with a relationship as a bag and assuming that at least one sentence in the bag can express the relationship between the two entities, thereby effectively reducing the impact of noisy data on distant supervision. Surdeanu et al. [11] proposed a multi-instance multi-label model based on probabilistic graphical models, which not only models noisy training data but also performs multi-class classification for entity pairs and their relationships. Experimental results demonstrated significant improvements in relation extraction performance. Liu et al. [12] proposed a weakly supervised relation extraction method based on convolutional neural networks for specific domains, constructing a semi-automatic pattern extraction system (EARES) to generate training corpora, which were then converted into vector feature matrices for classification model training using CNNs. Santos et al. [13] introduced a deep neural network approach for relation classification without manual features, but this sentence-based classifier could not be applied to large-scale knowledge bases due to the lack of manually annotated training data. Zeng et al. [14] combined multi-instance learning with neural network models to build a distant supervision-based relation extractor, achieving notable improvements though still far from satisfactory results.

Attention mechanisms, similar to human selective visual attention, aim to select information critical to the current task by computing probability distributions from numerous sources. Mnih et al. [15] incorporated attention mechanisms into RNN models for image classification, while Bahdanau et al. [16] applied similar attention mechanisms to machine translation tasks, performing translation and alignment simultaneously—marking the first application of attention mechanisms in NLP. Lin et al. [17] proposed a distant supervision relation extrac-

tion method using CNNs to extract sentence features, followed by constructing sentence-level attention mechanisms to address incorrect labeling issues.

2 GRU_ATT Relation Extraction Model

The overall model structure is shown in Figure 1 [Figure 1: see original paper]. First, all sentence sets containing the same entity pair are obtained and converted into vector representations. GRU is then used to extract semantic features, yielding sentence semantic feature vectors. Next, attention mechanisms calculate corresponding weights, and finally, the sentence semantic feature vectors are multiplied by their weights and summed to obtain the vector representation of set S.

2.1 Vectorization

2.1.1 Word Vectorization Given a sentence s consisting of t words {w , w , ···, w }, word2vec [18] maps each word w to a low-dimensional real-valued vector space. The vectors of all words in the sentence are concatenated to form the sentence vector. Word vectorization is performed using Equation (1):

$$x_i = V \cdot e_i$$

where e_i is the one-hot representation of word w_i, $V \in \mathbb{R}^{d_w \times m}$ is the word vector matrix, m is a fixed-size vocabulary, and d_w represents the word vector dimension. This yields the vectorized representation of each word in the sentence.

2.1.2 Position Vectorization In relation extraction tasks, words near entities often better highlight the relationship between the two entities in a sentence. Therefore, to more accurately express sentence meaning, the distance from each word in the sentence to both entities is concatenated into the word' s vector representation. If the word vectorization dimension is d_w and the position vectorization dimension is d_p, the sentence vector dimension becomes $d = d_w + 2d_p$.

For example, in the sentence "Beijing is the capital of China" shown in Figure 2 [Figure 2: see original paper], with "Beijing" and "China" as the two entities, the word "is" has a distance of -1 to "Beijing" and 4 to "China" , while "the" has distances of -2 and 3 respectively.

2.2 GRU Construction

LSTM (Long Short-Term Memory) is an improved RNN capable of learning long-term dependencies, proposed by Hochreiter & Schmidhuber in 1997 and later refined by Alex Graves. GRU [19] is a variant of LSTM that maintains LSTM' s effectiveness while simplifying the structure, reducing training parameters, and improving training speed. LSTM contains three gate computations: forget gate, input gate, and output gate, which add or remove information from the

cell state. GRU combines the forget and input gates into a single update gate while mixing cell states and hidden states. The update gate controls whether information from the previous time step is brought into the current cell state. The GRU structure is shown in Figure 3 [Figure 3: see original paper].

Using the vectorized sentence representation from the previous section as input, we illustrate the feature values of each state in the GRU unit for the i-th word:

$$z_t = \text{rec}(W_z \cdot x_t + U_z \cdot h_{t-1} + b_z)$$
$$r_t = \text{rec}(W_r \cdot x_t + U_r \cdot h_{t-1} + b_r)$$
$$\tilde{h}_t = \tanh(W_h \cdot x_t + r_t \odot U_h \cdot h_{t-1} + b_h)$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

where the rec activation function uses ReLU, x_t represents the current time step input, h denotes the memory cell, W represents weight matrices, and b denotes bias terms.

2.3 Attention Mechanism Introduction

Assume $S = \{s_1, s_2, ..., s_n\}$ is the set of all sentences containing entity pair $\langle e_1, e_2 \rangle$. To fully utilize information from each sentence in the bag, we introduce an attention mechanism to compute attention probabilities that reflect the importance of each sentence to the bag. When determining whether entity pair $\langle e_1, e_2 \rangle$ has relationship vector r, each sentence s_i in set S contains information about whether it expresses relationship r. First, set S is converted to vector form v using:

$$v = \sum_{i=1}^{n} \beta_i x_i$$

where β_i represents the weight of sentence s_i. To verify the impact of introducing the attention mechanism, we define β_i in two ways: First, by setting $\beta_i = 1/n$, which assumes all sentences in the bag are equally important for expressing relationship r. This is clearly unreasonable since some sentences do not express relationship r, and such noisy data negatively impacts results. Second, to avoid this issue and reduce the influence of noisy data, we accurately assign weights to each sentence by defining a function about x_i and r:

$$\text{score}(s_i, r) = x_i^T A r$$

This function describes the matching degree between sentence s_i and predicted relationship r, with value range [0,1], where 0 indicates s_i cannot express relationship r, and 1 indicates it definitely expresses r. A is a diagonal matrix.

The weight of the i-th sentence in the bag is then obtained through the softmax function:

$$\beta_i = \frac{\exp(\text{score}(s_i, r))}{\sum_{k=1}^{n} \exp(\text{score}(s_k, r))}$$

Substituting Equation (9) into Equation (7) yields the vector representation of set S:

$$v = \sum_{i=1}^{n} \beta_i x_i$$

After obtaining v, a linear function is defined to calculate scores for each possible relationship r:

$$y = M \cdot v + b$$

where M is a relation matrix and b is a bias term.

2.4 Model Training and Optimization

Model training employs minimization of the negative log-likelihood function. The conditional probability is defined through a softmax layer as:

$$P(r|S, \theta) = \frac{\exp(y_r)}{\sum_{j=1}^{n_r} \exp(y_j)}$$

where θ represents model training parameters with randomly initialized values, and n_r denotes the number of possible relationships. The optimization objective function $J(\theta)$ is then defined as:

$$J(\theta) = -\frac{1}{D} \sum_{i=1}^{D} \log P(r_i|S_i, \theta)$$

where D represents the number of training samples. Stochastic gradient descent is used to minimize this negative log-likelihood function.

3 Experimental Results

3.1 Experimental Data and Environment

The dataset was generated by aligning entity pairs from Freebase with the New York Times corpus (NYT). This dataset [10] was first developed by Riedel in 2010 and subsequently used by Raphael Hoffmann, Mihai Surdeanu, Lin, and

others. The dataset uses Stanford University' s named entity recognition tool to annotate the New York Times corpus, which is then matched with entities in Freebase. The dataset format is shown in Table 1 . The first column contains entity IDs, the second column contains the entities themselves, the third column shows the relationship, and the fourth column displays the sentence. For example, the triple (Hunan, contains, Changsha) from Freebase corresponds to the NYT sentence "one reason is that hunan' s fast-growing provincial capital changsha is beginning to siphon some workers," creating a data entry when the entities are matched.

The dataset contains 53 relationships (including the special NA relationship indicating no relationship between entities), 39,528 entities, 522,611 sentences in the training set (28,127 entity pairs, 18,252 relational facts), and 172,448 sentences in the test set (96,678 entity pairs, 1,950 relational facts).

Experimental Environment: Operating system Windows 7 64-bit; Processor Intel(R) CoreTM i5-4690; 8 GB RAM; Programming platform Pycharm, Python 2.7.

3.2 Evaluation Metrics

Precision and recall are collected to plot PR curves as evaluation metrics, calculated as follows:

$$\text{precision} = \frac{\text{right\_num}}{\text{out}}$$

$$\text{recall} = \frac{\text{right\_num}}{\text{all}}$$

where right_num represents correctly predicted instances, out represents total predictions, and all represents total test instances. Higher recall is better at the same precision level, and higher precision is better at the same recall level. Therefore, PR curves closer to the upper-right corner indicate better performance.

3.3 Parameter Settings

Softmax is used as the classifier. L2 regularization constrains network parameters, dropout is employed during training, and the Adadelta optimization method is used for model training. Specific parameter settings are shown in Table 2 .

3.4 Experimental Results and Analysis

3.4.1 Comparison with Existing Methods First, GRU_ATT is compared with traditional feature-based relation extraction methods, as shown in Figure 4 [Figure 4: see original paper]. The red curve represents GRU_ATT results, the

green curve shows Mintz' s [9] traditional distant supervision model from 2009, and the black curve displays MultiR [20], Hoffmann' s 2011 multi-instance learning model for handling overlapping relations. GRU_ATT outperforms both Mintz and MultiR across the entire recall range. While Mintz and MultiR decline rapidly when recall exceeds approximately 0.2, GRU_ATT remains relatively stable throughout, indicating that manually designed features cannot accurately express sentence semantics.

To demonstrate GRU' s advantages, we compare it with Lin et al.' s 2015 CNN_ATT method [17], which uses CNNs for sentence feature extraction followed by sentence-level attention mechanisms to address incorrect labeling. The comparison results are shown in Figure 5 [Figure 5: see original paper] (black: CNN_ATT; red: GRU_ATT). CNNs can only extract position-invariant features and cannot learn temporal sequences. In contrast, GRU' s memory modules can fully utilize entire sentence sequence information, including associations between words, making it more suitable for NLP tasks. The results show GRU_ATT outperforms CNN_ATT.

3.4.2 Impact of Attention Mechanism This experiment verifies the impact of introducing the attention mechanism. Three models are established: (1) GRU only without attention, (2) GRU_AVE where all sentences for an entity pair have equal weight, and (3) GRU_ATT. The comparison results are shown in Figure 6 [Figure 6: see original paper] (red: GRU_ATT; black: GRU_AVE; green: GRU). GRU_AVE, which introduces attention and considers sentence meaning, reduces noisy data impact and outperforms GRU. However, GRU_AVE treats all sentences equally, still introducing some noisy data from sentences expressing incorrect relationships. GRU_ATT achieves the highest precision across the entire recall range, demonstrating that attention mechanisms effectively reduce the impact of noisy data in distant supervision.

4 Conclusion

This paper proposes a novel neural network model GRU_ATT for textual relation extraction. The model overcomes long-distance dependency limitations of traditional deep learning models, significantly reduces time and effort for manual data labeling, and minimizes noisy data impact on results. Experiments on public corpora show improved PR curves. Although attention mechanisms reduce noisy data impact, the problem is not completely solved. Future work will focus on automatically discovering relationships in open domains for relation extraction.

References

[1] Zhao Yanyan, Qin Bing, Che Wanxiang, et al. Research on Chinese Event Extraction [J]. Journal of Chinese Information Processing, 2008, 22(1): 3-8.

[2] Chen Yu, Zheng Dequan, Zhao Tiejun. Chinese Relation Extraction Based on Deep Belief Nets [J]. Journal of Software, 2012, 23(10): 2572-2585.

[3] Huang Xun, You Hongliang, Yu Yang. A Review of Relation Extraction [J]. New Technology of Library and Information Service, 2013(11): 30-39.

[4] Kurt B, Colin E, Praveen P, et al. Freebase: a collaboratively created graph database for structuring human knowledge [C]// Proc of KDD. 2008: 1247-1250.

[5] Lehmann J. DBpedia: a nucleus for a web of open data [C]// Proc of Semantic Web, International Semantic Web Conference, Asian Semantic Web Conference. 2007: 11-15.

[6] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735.

[7] Rink B, Harabagiu S. UTD: classifying semantic relations by combining lexical and semantic resources [C]// Proc of International Workshop on Semantic Evaluation. Association for Computational Linguistics. 2010: 144-149.

[8] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations [C]// Interactive Poster and Demonstration Sessions. Stroudsburg: Association for Computational Linguistics, 2004: 22.

[9] Mintz M, Steven B, Rion S, et al. Distant supervision for relation extraction without labeled data [C]// Proc of Joint Conference of the Meeting of the ACL and the International Joint Conference on Natural Language Processing of the Afnlp: Volume. Stroudsburg: Association for Computational Linguistics, 2009: 1003-1011.

[10] Riedel S, Yao Limin, Mccallum A. Modeling relations and their mentions without labeled text [C]// Proc of European Conference on Machine Learning and Knowledge Discovery in Databases. Springer-Verlag, 2010: 148-163.

[11] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction [C]// Proc of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2010: 455-465.

[12] Liu Kai, Fu Haidong, Zou Yuwei, et al. Chinese medical weak supervised relation extraction based on convolution neural network [J]. Computer Science, 2017, 44(10): 249-253.

[13] Santos C N D, Gattit M. Deep convolutional neural networks for sentiment analysis of short texts [C]// Proc of International Conference on Computational Linguistics. 2014.

[14] Zeng Daojian, Liu Kang, Chen Yubo, et al. Distant supervision for relation extraction via piecewise convolutional neural networks [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2015: 1753-1762.

[15] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention [EB/OL]. (2014-12-03). https://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention.pdf.

[16] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. Computer Science, 2014.

[17] Lin Yankai, Shen Shiqi, Liu Zhiyuan, et al. Neural relation extraction with selective attention over instances [C]// Proc of Meeting of the Association for Computational Linguistics. 2016: 2124-2133.

[18] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality [EB/OL]. (2014-11-25). https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[19] Dey R, Salemt F M. Gate-variants of Gated Recurrent Unit (GRU) neural networks [C]// Proc of IEEE International Midwest Symposium on Circuits and Systems. 2017: 1597-1600.

[20] Hoffmann R, Zhang Congle, Ling Xiao, et al. Knowledge-based weak supervision for information extraction of overlapping relations [C]// Proc of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2011: 541-550.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.