Relation Extraction with Neural Networks

Golnar Sheikhshab

1 Task

The objective is to extract relations between pre-detected named entities. For this project, I will limit the relations to be binary (where there are exatcly two name entities related to each other) and the relation sets to be ones with signle members (the output is binary: either the relation of interest exists or it doesn't).

2 Dataset

I am planning to work on the corpus of event extraction for the Bacteria Biotope in BioNLP shared task 2016. A train set, a dev set, and a test set are publicly available at the shared task webpage. Each of these datasets contain several sets of files, each set for one abstract. In each set, there is a file for the raw text; one file with the named entities categorized as BACTERIA, HABITAT, or GEOGRAPHICAL; and one file containing all LIVES_IN relation. There are exactly two named entities participating in each relation: the first is in category BACTERIA, and the second one in either of HABITAT or GEOGRAPHICAL categories. Figure 1 shows an example abstract with pre-detected entities and relationships between some of those entities.

To develop a broad understanding of the causes and patterns of occurrence of wheezing associated respiratory infections, we analyzed data from an 11-year study of acute lower respiratory illness in a pediatric practice. Although half of the WARI occurred in children less than 2 years of age, wheezing continued to be observed in 19% of children greater than 9 years of age who had lower respiratory illness. Males experienced LRI 1.25 times more often than did females; the relative risk of males for WARI was 1.35. A nonbacterial pathogen was recovered from 21% of patients with WARI; respiratory syncytial virus, parainfluenza virus types 1 and 3, adenoviruses, and Mycoplasma pneumoniae accounted for 81% of the isolates. Patient age influenced the pattern of recovery of these agents. The most common cause of WARI in children under 5 years of age was RSV whereas Mycoplasma pneumoniae was the most frequent isolate from school age children with wheezing illness. The data expand our understanding of the causes of WARI and are useful to diagnosticians and to researchers interested in the control of lower respiratory disease.

Figure 1: example abstract with pre-detected BACTERIA and HABITAT entities. Lives_In relations are shown using colored underlines.

3 Approach

I intend to learn Neural Net based dense representations for sentences and named entities based on a large collection of text (for example PubMed abstracts) and add a classification layer to the architecture. We can then try a transductive graph propagation step. The nodes of the graph used in graph-propagation will be the dense representations we learnt.

I intend to test three different representations: 1) representation of the whole sentence only; 2) representation of the whole sentence concatenated to representations between the two entities; 3) representation of the sentence where the entity mentions were removed concatenated to representations of the entities.

3.1 How to get dense representations

There are two types of representations involved in my approach: entity representations and sentence representations. I intend to use a continuous bag of words (CBOW) approach for entity representations where the skip-gram word embeddings [Mikolov *et al.*, 2013] are averaged to give an embedding for the entity. The idea in skip-gram is to predict the surrounding words given the representation of the current word. For the sentence representations, I intend to follow the idea of skip-thoughts [Kiros *et al.*, 2015], the counterpart of skip-grams for sentences where gated recurrent neural nets [Cho *et al.*, 2014] are used to predict surrounding sentences given the representation of a sentence.

3.1.1 Entity Embeddings

Entity embeddings will be an average of word embeddings obtained by skip-gram. In skip-gram (illustrated in Figure 2), we predict surrounding words based on the current word. The current word and candidates for the surrounding words are the input to the model and the binary output indicates if the candidates are correct. Also, to tame the time complexity of objective optimization, only a small sample of vocabulary is used as negative examples and the model is trained to discriminate between correct candidates and sampled noise. Therefore, if we are predicting the next word based on the current word, the objective is

$$\sum_{t} \Big[log(Q(w_t, w_{t+1})) + \sum_{w' \in sampled noise} log(1 - Q(w_t, w')) \Big] \tag{1}$$

where Q is the logistic regression of dot product between the embeddings of the current word and the candidate output word:

$$Q(w_1, w_2) = \frac{exp((w_1 W_h).(w_2 W_h))}{1 + exp((w_1 W_h).(w_2 W_h))}. \tag{2}$$

The parameters of the model are items of W_h matrix that is in fact the matrix of embeddings. all w's are one-hot vectors of the words.

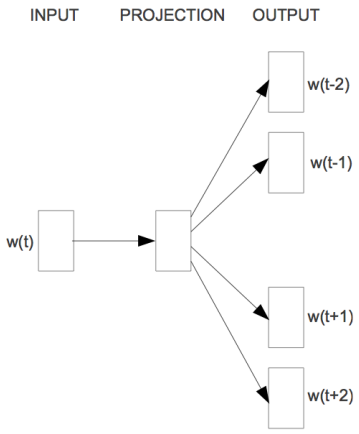

Figure 2: skip-gram

3.1.2 Sentence Embeddings

skip-thought vectors are obtained by encoding a sentence using a gated recurrent neural network and decoding two surrounding sentences using two other gated recurrent neural networks where the objective is to maximize the probability of observed data.

Equations 3 to 6 are the update equations for encoding sentence s_i. After encoding finishes we end up with h_i, the final hidden representation of s_i that is used in decoding.

$$r_i^t = \sigma(W_r x_i^t + U_r h_i^{t-1}) \tag{3}$$

$$z_i^t = \sigma(W_z x_i^t + U_z h_i^{t-1}) \tag{4}$$

$$\bar{h}_i^t = tanh(W x_i^t + U(r_i^t \odot h_i^{t-1}) \tag{5}$$

$$h_i^t = (1 - z_i^t) \odot h_i^{t-1} + z_i^t \odot \bar{h}_i^t \tag{6}$$

Equations 7 to 10 and equations 11 to 14 are update equations in decoders of next and previous sentences. These equations are similar to encoding equations except for presence of C, C_r, and C_z bias matrices.

$$r_{i+1}^t = \sigma(W_r^{d1} x_{i+1}^{t-1} + U_r^{d1} h_{i+1}^{t-1} + C_r^{d1} h_i) \tag{7}$$

$$z_{i+1}^t = \sigma(W_z^{d1} x_{i+1}^{t-1} + U_z^{d1} h_{i+1}^{t-1} + C_z^{d1} h_i) \tag{8}$$

$$\bar{h}_{i+1}^t = tanh(W^{d1} x_{i+1}^{t-1} + U^{d1}(r_{i+1}^t \odot h_{i+1}^{t-1} + C^{d1} h_i) \tag{9}$$

$$h_{i+1}^t = (1 - z_{i+1}^t) \odot h_{i+1}^{t-1} + z_{i+1}^t \odot \bar{h}_{i+1}^t \tag{10}$$

$$r_{i-1}^t = \sigma(W_r^{d2} x_{i-1}^{t-1} + U_r^{d2} h_{i-1}^{t-1} + C_r^{d2} h_i) \tag{11}$$

$$z_{i-1}^t = \sigma(W_z^{d2} x_{i-1}^{t-1} + U_z^{d2} h_{i-1}^{t-1} + C_z^{d2} h_i) \tag{12}$$

$$\bar{h}_{i-1}^t = tanh(W^{d2} x_{i-1}^{t-1} + U^{d2}(r_{i-1}^t \odot h_{i-1}^{t-1} + C^{d2} h_i) \tag{13}$$

$$h_{i-1}^t = (1 - z_{i+1}^t) \odot h_{i+1}^{t-1} + z_{i+1}^t \odot \bar{h}_{i+1}^t \tag{14}$$

Finally, equation 15 presents the objective of the system to be maximized.

$$objective = \sum_i \Big[\sum_t logP(w_{i+1}^t | w_{i+1}^{<t}, h_i) + \sum_t logP(w_{i-1}^t | w_{i-1}^{<t}, h_i) \Big] \tag{15}$$

where

$$P(w_{i+1}^t | w_{i+1}^{<t}, h_i) \propto exp(v_{w_{i+1}^t} . h_{i+1}^t) \tag{16}$$

$$P(w_{i-1}^t | w_{i-1}^{<t}, h_i) \propto exp(v_{w_{i-1}^t} . h_{i-1}^t) \tag{17}$$

Figure 3 shows the encoder-decoder model in skip-thoughts.

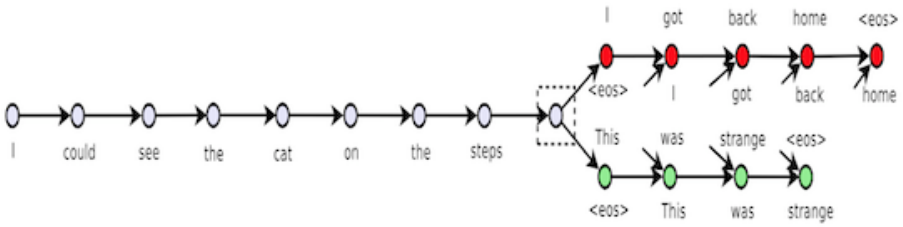

Figure 3: skip-thoughts

References

[Cho *et al.*, 2014] Cho, Kyunghyun, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. "On the properties of neural machine translation: Encoder-decoder approaches." arXiv preprint arXiv:1409.1259 (2014).

[Kiros *et al.*, 2015] Kiros, Ryan, Yukun Zhu, Ruslan R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. "Skip-thought vectors." In Advances in neural information processing systems, pp. 3294-3302. 2015.

[Mikolov *et al.*, 2013] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).