Two Papers of Distant Supervision for Relation Extraction

huweilong@whu.edu.cn

武汉大学计算机学院

2017 年 10 月 13 日

Table of contents

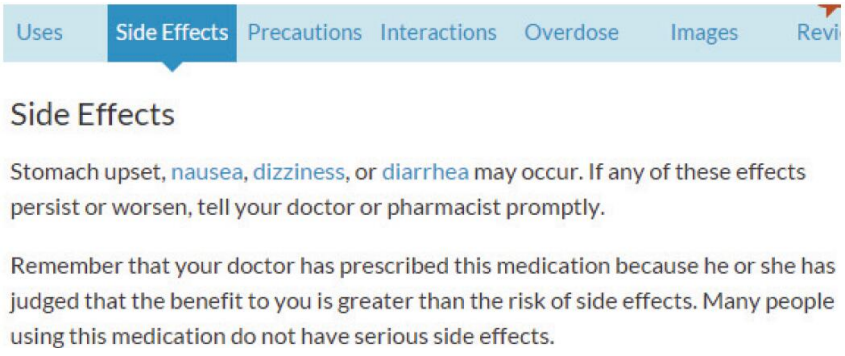
Distant supervision is less expensive to obtain than directly supervised labels,but produces noisy training data

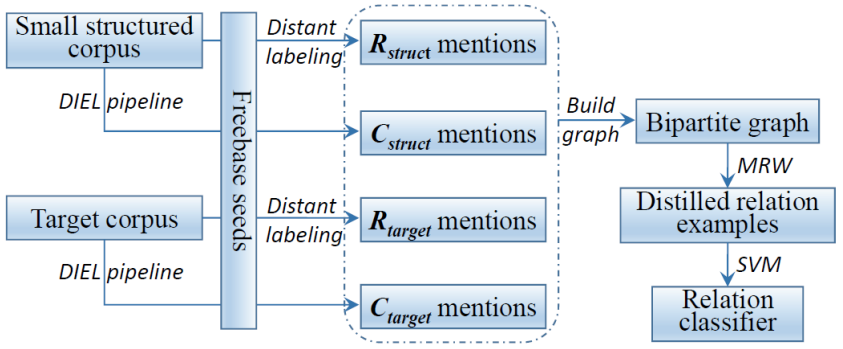
- Distant supervision is often coupled with learning methods that allow for noise
- Combine distant labeling with label propagation(Bing et al.2015;2016)

This paper presents three new contributions

1. we combine the concept-instance extraction and relation extraction tasks
2. a novel use of document strcture,some small well-structured corpora
3. perform extensive experiments, and show substantial improvements

DIEJOB:Distant IE by JOint Bootstrapping





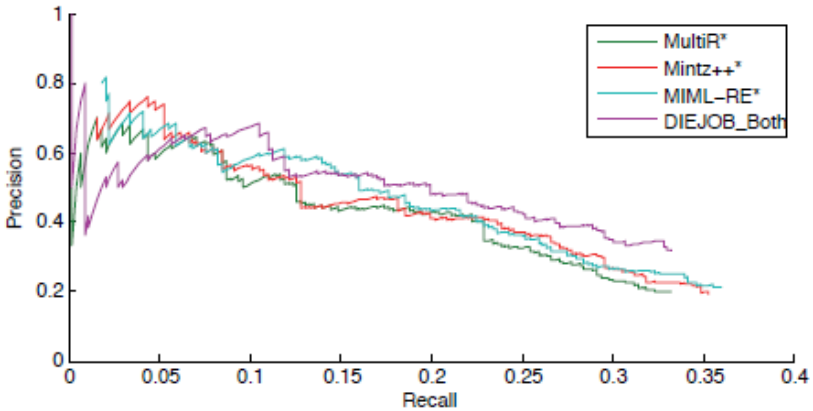
- From each corpus, DIEJOB produces two types of metion sets, relation mention set R and concept metion set C
- 3 drug relations and 2 corresponding concept types, 5 disease relations and 5 corresponding concept types
- Extract concept instances from Freebase as seeds, and extend the seed set using LP
- Use multi-class label propagation method **MultiRankWalk**

Experiments

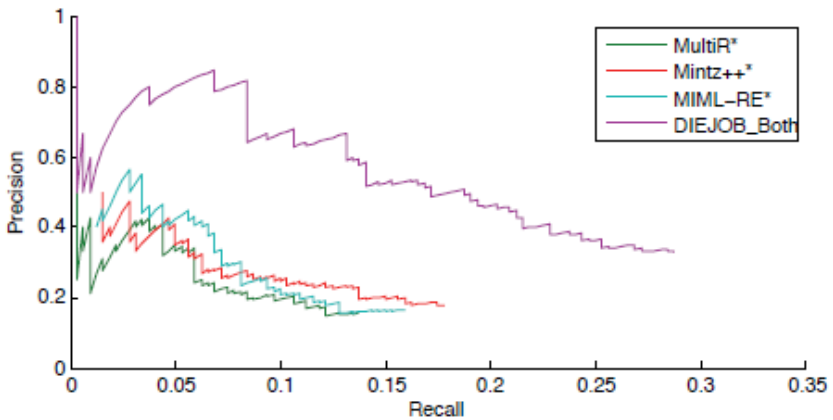
Corporas or Dataset

	target	structured
disease	WikiDisease	MayoClinic
drug	DailyMed	WebMD

	Disease			Drug		
	P	R	F1	P	R	F1
DS_Struct	0.300	0.300	0.300	0.232	0.072	0.110
DS_Target	0.228	0.335	0.271	0.170	0.188	0.178
DS_Both	0.233	0.353	0.281	0.154	0.175	0.164
DIEBOLDS	0.143	0.372	0.209	0.050	0.435	0.090
MultiR*	0.198	0.333	0.249	0.156	0.138	0.146
Mintz++*	0.192	0.353	0.249	0.177	0.178	0.178
MIML-RE*	0.211	0.360	0.266	0.167	0.160	0.163
DIEJOB_Target	0.231	0.337	0.275	0.299	0.300	0.300
DIEJOB_Both	0.317	0.333	**0.324**	0.327	0.288	**0.306**
DIEJOB_Target*	0.235	0.339	0.277	0.289	0.425	0.344
DIEJOB_Both*	0.317	0.333	0.324	0.282	0.422	0.338



(a) Disease domain.



(b) Drug domain.

- (Riedel, Yao, and McCallum 2010)

  "at least one" heuristic.

- MutilR(Hoffmann et al.2011), MIML-RE(Surdeanu et al.2012)

  support multiple relations expressed in a bag.

- (Bing et al.2016)

  use document structure to enrich LP graph

- (Riloff and Jones 1999; Agichtein and Gravano 2000; Bunescu and Mooney 2007)

  classic bootstrap learning scheme

Table of contents

Introduction

Freebase /location/location/contains (Nevada, Las Vegas)

S1. **[Nevada] then sanctioned the sport , and the U.F.C. held its first show in [Las Vegas] in September 2001.**

S2. Pinnacle owns casinos in [Nevada], Louisiana , Indiana , Argentina and the Bahamas , but not in the top two American casino cities , Atlantic City and [Las Vegas].

S3. **He has retained two of [Nevada] 's most prominent criminal defense lawyers , Scott Freeman of Reno and David Chesnoff of [Las Vegas].**

S4. The state 's population is growing , but not skyrocketing the way it is in Arizona and [Nevada] , and with no city larger than 100,000 residents , Montana essentially does not have suburbs or exurbs like those spreading around Phoenix, [Las Vegas] and Denver.

Descriptions

[Nevada]: Nevada is a state in the Western, Mountain West, and Southwestern regions of the United States.

[Las Vegas]: officially the City of Las Vegas and often known as simply Vegas, is a city in the United States, the most populous city in the state of Nevada, the county seat of Clark County, and the city proper of the Las Vegas Valley.

Training instances of the triplet */location/location/contains (Nevada, Las Vegas)*

1. A bag may contain multiple valid sentences

   consider multiple valid sentences and extract features by neural networks

2. The entity descriptions, which can provide helpful background knowledge

   none of the existing work uses them for RE under distant supervision

main work

Propose a sentence-level attention model-APCNNs,which extracts sentence features using PCNNs and learns the weights of sentences by the attention module. use convolutional neural networks to extract entity descriptions' feature vectors-APCNNs+D.

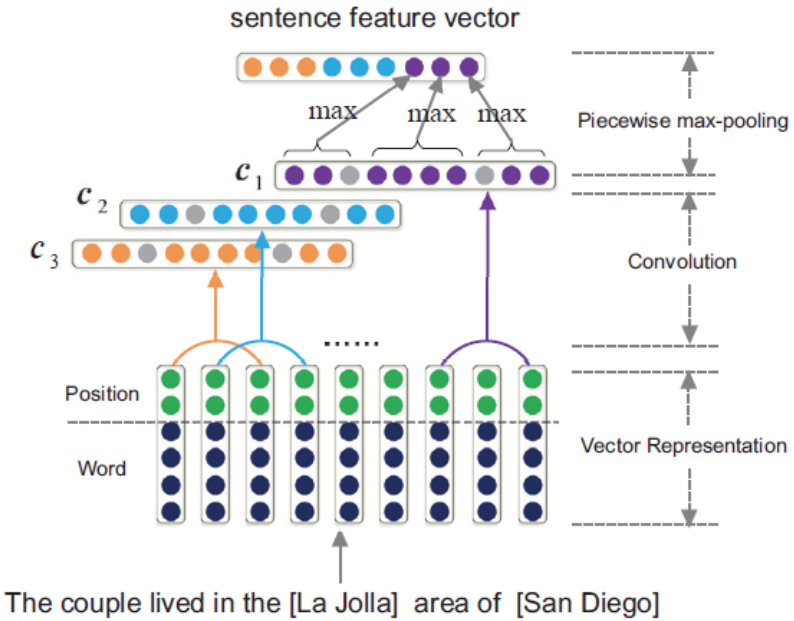
main contributions:

1. introduce a sentence-level attention model to select multiple valid sentences in a bag
2. use entity descriptions to provide background knowledge for predicting relations and entity representations
3. achieve state-of-the-art preformance

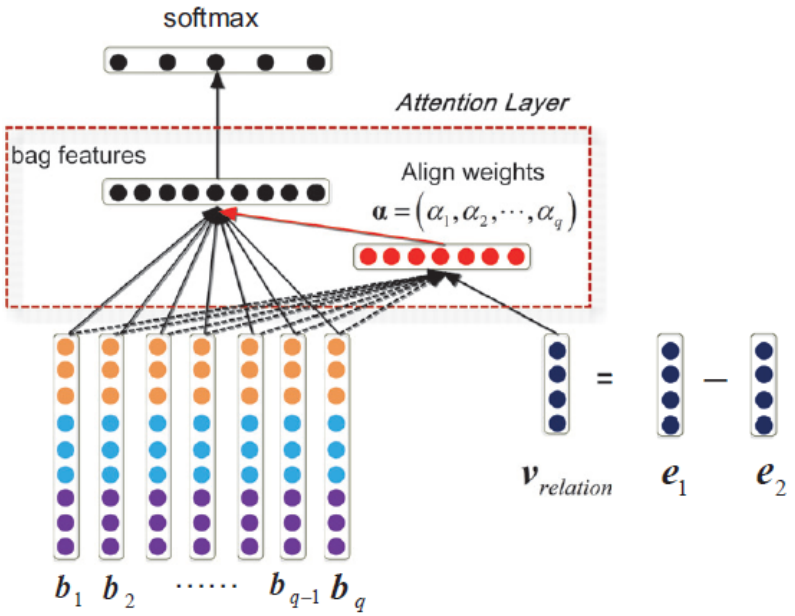
Suppose that there are N bags $\{B_1, B_2, \cdots, B_N\}$ in the training set, the i-th bag contains instances $B_i = \{b_1^i, b_2^i, \cdots, b_{q_i}^i\}(i = 1, 2, \cdots, N)$

task

The objective of multi-instance learning is to predict the labels of the unseen bags. We need to learn a relation extractor based on the training data and then use it to predict relations for test set.



(a) PCNNs Module



(b) Sentence-level Attention Module

Neural network architecture of APCNNs

- PCNNs Module: Vector Representation, Convolution and Piecewise Max-pooling
- Sentence-level Attention: Module Attention Layer and Softmax Classifier

Word Embeddings

Employ the method (Mikolov et al. 2013a) to train word embeddings and denote it by E.

Position Embeddings

the size of word representation is k_w and that of position representation is k_d, then the size of a word vector is $k = k_w + 2k_d$.

Convolution

n weight matrices $\widehat{W} = \{W_1, W_2, \cdots, W_n\}$,all the convolution operations can be expressed by

$$c_{ij} = W_i \otimes S_{(j-w+1):j}$$

where $1 \leq i \leq n \quad and \quad 1 \leq j \leq |S| - w + 1$

Picewise Max-pooling

In order to capture the structural information and fine-grained features, PCNNs divides an instance into three segments according to the given entity pair (two entities cut the sentence into three parts) and do max-pooling operation on each segment.

Attention Layer

Specifically, for a bag labeled by $r(e_1, e_2)$, vector $v_{relation} = e_1 - e_2$

$$\alpha_i = \frac{exp(\omega_i)}{\sum_{j=1}^{q} exp(\omega_j)} \quad \omega_i = W_a^T(tanh[b_i; v_{relation}]) + b_a$$

bag feature:

$$\bar{b} = \sum_{i=1}^{q} \alpha_i b_i$$

Softmax

$$o = W_s \bar{b} + b_s \quad p(r_i|B;\theta) = \frac{exp(o_i)}{\sum_{j=1}^{n_o} exp(o_j)}$$

Let $\theta = (E, \widehat{W}, PF_1, PF_2, W_a, W_s)$

Use another traditional CNN (a convolution layer and a single max-pooling layer) to extract features from entity descriptions. $\mathcal{D} = \{(e_i, d_i)|i = 1, 2, \cdots, |\mathcal{D}|\}$. Loss Function:

$$\mathcal{L}_e = \sum_{i=1}^{|\mathcal{D}|} ||e_i - d_i||_2^2$$

Extract descriptions for entities from Freebase and Wikipedia pages.extract the first 80 words for descriptions

- Train the APCNNs which only contains the sentence-level attention module (no entity descriptions). Then we define the objective function using cross-entropy as follows.

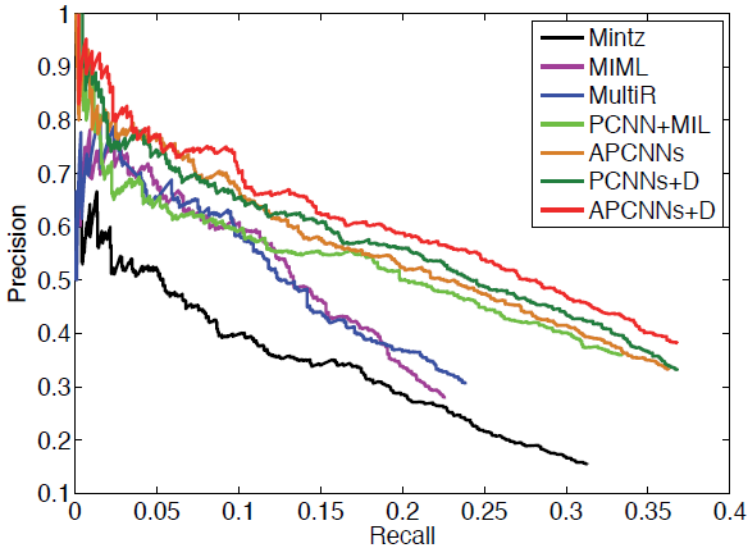
$$min\mathcal{L}_A = \sum_{i=1}^{N} logp(r_i|B_i, \theta)$$

- Train the model APCNNs+D which contains both sentence-level attention module and entity descriptions. The objective function is

$$min\mathcal{L} = \mathcal{L}_A + \lambda\mathcal{L}_e$$

Dataset: New York Times(NYT)

Evaluation Metrics: Held-out evaluation and Human evaluation



Accuracy (%)	Top 100	Top 200	Top 500	Average
Mintz	0.77	0.71	0.55	0.676
MultiR	0.83	0.74	0.59	0.720
MIML	0.85	0.75	0.61	0.737
PCNNs+MIL	0.86	0.80	0.69	0.783
APCNNs	**0.87**	0.82	0.72	0.803
PCNNs+MIL+D	0.86	0.82	0.71	0.797
APCNNs+D	**0.87**	**0.83**	**0.74**	**0.813**

Bag Label	Instances	APCNNs	APCNNs+D
/location/location/contains (New Orleans, Dillard University)	1. She graduated from [Dillard University] in [New Orleans] and received a master's degree in marine science from the College of William and Mary.	0.223	0.239
	2. Jinx Broussard, a communications professor at [Dillard University] in [New Orleans], said four members of her family had lost their houses to the hurricanes.	0.216	0.235
	3. I was grieving from the death when I graduated from high school, but I decided to go to [Dillard University] in [New Orleans].	0.268	0.242
	4. When he came here in May 2003 to pick up an honorary degree from [Dillard University], his dense schedule didn't stop him from calling Dooky Chase's, the Creole restaurant he sang about in "Early in the Morning Blues, "where he'd eaten his favorite dish ever since he lived in [New Orleans] in the 1950's."	0.090	0.073
	5. He is waiting because of his involvement with a group that oversees a partnership between the University of Colorado and [Dillard University] in [New Orleans]."	0.203	0.211

Related Work

1. Supervised Methods
   - GuoDong et al., (2005)
     explored a set of feature(lexical and syntactic)
   - Bunescu and Mooney 2005; Zelenko, Aone and Richardella, (2003)
     used kernel methods(subsequence kernel and dependency tree kernel)
   - Zeng et al., (2014)
     exploited a CNN to extract lexical and sentence level feature
   - dos Santos, Xiang and Zhou, (2015)
     proposed a Classification by Ranking CNN(CR-CNN) model

2. Distant Supervised Methods
   - Mintz et al., (2009)
     extracted features from all sentences
   - Riedel, Yao, and McCallum,(2010), Hoffmann et al., (2011) and Surdeanu et al., (2012)
     used graphical model to select the valid sentences
   - Nguyen and Moschitti, (2011)
     utilized relation definations and Wikipedia documents
   - Zeng et al., (2015)
     used PCNNs to automatically learn sentence level features
   - Lin et al., (2016)
     proposed to use attention to select informative sentences

Thank You

huweilong@whu.edu.cn