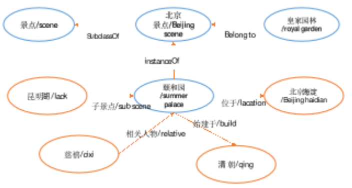
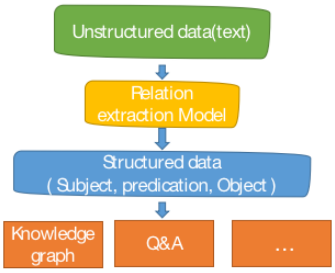
Type-aware Open Information Extraction via Graph Augmentation Model

Qinghua Wen, Yunzhe Tian, Lei Hou Juanzi Li
Department of Computer Science and Technology, Tsinghua University, China 100084

◆ Motivation and Challenge

■ Motivation




Unstructured data(text)

Relation extraction Model

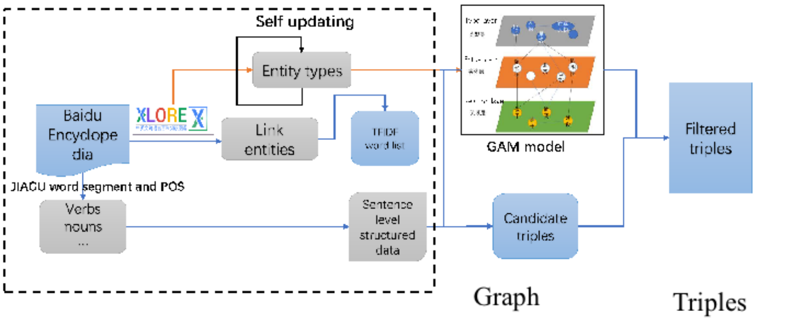
Structured data (Subject, predication, Object)

Knowledge graph | Q&A | …

■ Problem and Challenge

Few Chinese corpus.
Word segmentation and POS need improve.
Lacking evaluation and score for relation.

■ Method framework



Data preprocessing

Graph augmentation model

Triples scoring

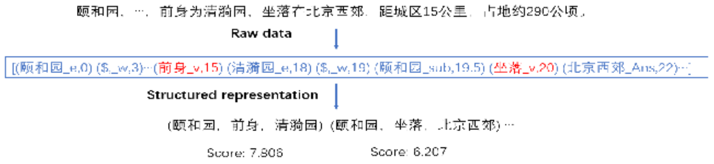
◆ Data Preprocessing

- **Raw data**: Baidu Baike texts.
- **Linked entities**: Owned by Baidu Baike texts and use the XLORE to link entity mentions in the texts.
- **Word segmentation and POS**: *jiagu* NLP tool with the entity dictionary.
- **rule-based candidate triple extraction method:**

1) Verbs in each sentence as relation words.

2) The head entities and tail entities according to their relative position of the relation words in the sentence.

3) filter entities with TFIDF, and finally get the triples.


颐和园，…，前身为清漪园，坐落在北京西郊，距城区15公里，占地约290公顷。

Raw data

[(颐和园_e,0) ($,_w,3)…(前身_v,15) (清漪园_e,18) ($,_w,19) (颐和园_sub,19.5) (坐落_v,20) (北京西郊_Ans,22)…]

Structured representation

(颐和园，前身，清漪园) (颐和园，坐落，北京西郊) …

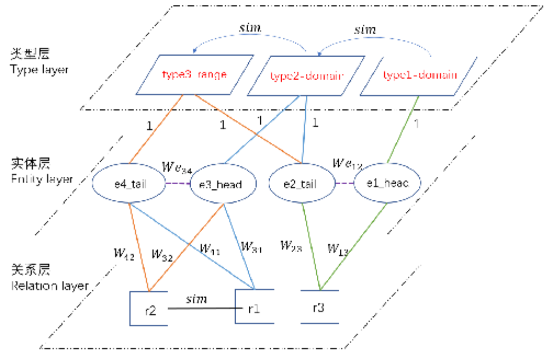
Score: 7.806 Score: 6.207

■ Generation priority

Head entity: subject entity > linked entity > nearest noun.
Tail entity: added entity > linked entity > nearest noun.
Triples: {Head entity, relation, Tail entity}

◆ Graph Augmentation Model(GAM)

■ Three layers graph model



■ Layers construction

Relation layer Nodes: verbs.
Edges: nodes similarity >σ = 0.7.
Score: initialize the importance score of each relation to 1.

Entity layer Nodes: co-occur entity with relation.
Edges: W_{ij} Frequency between co-occur entity and relation.
Score: initialize to 1.

Type layer Nodes: 50 coarse-grained types.
Edges: PMI similarity of types.
Score: initialize to 1.

■ Importance propagation and Triple scoring

Hypothesis 1 The entities linked by many important relations and many important types tend to be important, the relations linked by many important entities tend to be important, and the types linked by many important entities tend to be important.

Hypothesis 2 The relations linked by many important relations tend to be important, and the types linked by many important types tend to be important.

$$s_1(e_i)^{k+1} = \sum_{\forall m: r_m - e_i} s(r_m)^k \frac{w(r_m - e_i)}{\sigma_{\forall n: r_m - e_n} w(r_m - e_n)} + \sum_{\forall m: t_m - e_i} s(t_m)^k \frac{w(t_m - e_i)}{\sigma_{\forall n: t_m - e_n} w(t_m - e_n)}$$

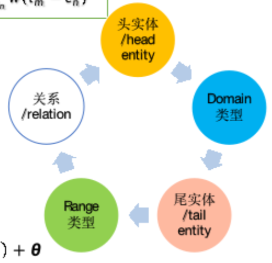
$$s_2(r_i)^{k+1} = (1-\beta) \times s(r_i)^k + \beta \times \sum_{\forall j: r_j - r_i} s(r_j)^k \frac{w(r_j - r_i)}{\sigma_{\forall n: r_j - r_n} w(r_j - r_n)}$$

$$s(e_i)^{k+1} = s_1(e_i)^{k+1}$$

$$s(r_j)^{k+1} = (1-a) \times s_1(r_j)^{k+1} + a \times s_2(r_j)^{k+1}$$

$$s(t_i)^{k+1} = (1-a) \times s_1(t_i)^{k+1} + a \times s_2(t_i)^{k+1}$$

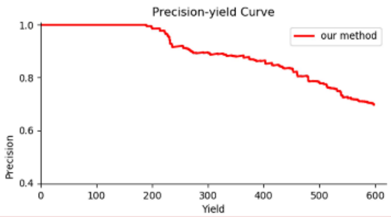
$$S_{tr} = lo\ (\alpha * (S_{do\ ma\ in} * S_{ra\ \ ge}) + (1-\alpha) * \beta * (S_{eh} + S_{re} + S_{et}) + \theta$$



◆ Experiments

- **DataSet:** 1,218 web pages related to Beijing attractions, 91,649 sentences and 12,932 entities.
- **Experimental results**

Method	No. of Triples	Precision	Yield
Our Method (threshold 4)	7726	**77%**	5949
Our Method (threshold=3)	**10154**	70%	**7107**
DSNFs [8]	9292	58%	5459
UnCORE [11]	2038	41.2%	841



Score	Triples	Documents
10.050	颐和园/坐落/北京西郊	颐和园
10.050	Summer Palace/located in/Beijing western suburbs	Summer Palace
9.079	荷花/成为/著名景观	什刹海
9.079	lotus/becomes/famous scenery	Shichahai
7.275	恭王府/占地/61120 平方米	恭王府
7.275	Prince kung's Mansion/covers/61120 square meters	Prince kung's Mansion
5.010	东南角楼/建于/明朝	明城墙遗址公园
5.010	Southeast turret/built in/Ming	Wall Ruins Park
4.487	嘉靖/重建/中极殿	故宫博物院
4.487	Jiajing/rebuilt/Zhongji Hall	Palace Museum

Qinghua Wen (wtsinghua1@163.com), 2020-11-04