Motivation and Challenge

- **Motivation**
  - Unstructured data (text)
  - Relation extraction model
  - Structured data (subject, predicate, object)

- **Problem and Challenge**
  - Few Chinese corpus.
  - Word segmentation and POS need improve.
  - Lacking evaluation and score for relation.

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 TF-IDF, and finally get the triples.

Graph Augmentation Model (GAM)

- **Three layers graph model**

- **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.

Experiments

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

  **Experimental results**

  ![Graph Augmentation Model](image_url)

  **Layers construction**
  - **Relation layer** Nodes: verbs.
  - Edges: nodes similarity $\geq 0.7$.
  - Score: initialize the importance score of each relation to 1.
  - **Entity layer** Nodes: co-occur entity with relation.
  - Edges: $\mathbf{W}_j$: 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.

  ![Graph Augmentation Model](image_url)

  **Final scoring**
  - $s_\alpha = b \cdot \left( s_\alpha \cdot s_\alpha \cdot s_\alpha \right) + \left( 1 - \alpha \right) \cdot f \cdot \left( s_\alpha + s_\beta + s_\gamma \right) + \theta$

  ![Graph Augmentation Model](image_url)

  **Data pre-processing**

  - **Raw data**: Baidu Baike texts.
  - **Link 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 TF-IDF, and finally get the triples.