# ICDM 2019 Knowledge Graph Contest: Team UWA

2019 IEEE International Conference on Data Mining (ICDM)
Michael Steart, Majigsuen Enkhsaikhan, Wei Liu

Zihao Li, 2020.03.22

# Introduction about ICDM/ICBK 2019 Contest

- **Practical engineering background:** Automobile Engineering, Catering Service, Cosmetics, Public Security;
- **Requirements:** Each competition team is invited to build a model that takes an article as input and outputs a graph;
- **Evaluation:** Team submissions will be judged by competition organizers on (a) their overall quality of the constructed knowledge graphs, and (b) generalization ability of their methodology in multiple domains;
- **Dataset:** The dataset consists of 300 recent published articles from news media of 4 industries. Each article is of 150 to 250 words, contains around 8-20 entities;

# Example

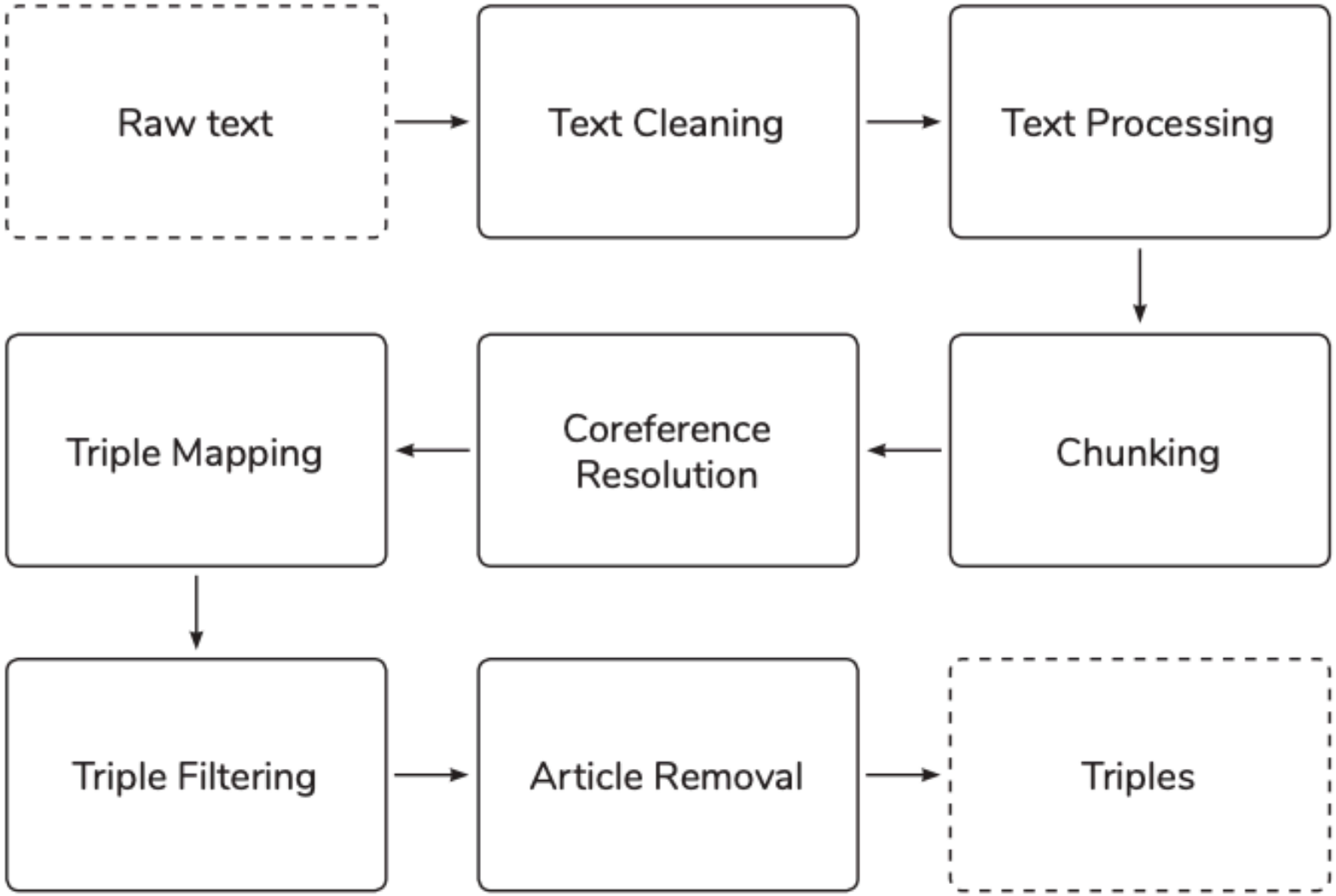
**Input:** BYD debuted its E-SEED GT concept car and Song Pro SUV alongside its all-new e-series models at the Shanghai International Automobile Industry Exhibition. The company also showcased its latest Dynasty series of vehicles, which were recently unveiled at the company's spring product launch in Beijing. A total of 23 new car models were exhibited at the event, held at Shanghai's National Convention and Exhibition Center, fully demonstrating the BYD New Architecture (BNA) design, the 3rd generation of Dual Mode technology, plus the e-platform framework. Today, China's new energy vehicles have entered the 'fast lane', ushering in an even larger market outbreak. Presently, we stand at the intersection of old and new kinetic energy conversion for mobility, but also a new starting point for high-quality development. To meet the arrival of complete electrification, BYD has formulated a series of strategies, and is well prepared.

# Example

## Output

# The Pipeline of Triple Extraction

- **Raw text:** Input;
- **Text Cleaning, Text Processing, Chunking:** delete noise, POS tagging, NER;
- **Coreference Resolution:** merging the same entity with different descriptions;
- **Triple Mapping, Triple Filtering and Article Removal:** Extract relations, remove stop words, unmeaning tokens;
- **Triples:** output



# Triple Extraction

1. **Text Cleaning:** Text data is cleaned to manage special characters such as hyphen and quotation marks and also break sentences joined together with no space between them;
2. **Text Processing:**

- **Tokenisation:** word segmentation based on blank or other specific characters;
- **POS tagging:** apply SpaCy tool kit to realize part-of-speech tagging;
- **NER:** recognize proper nouns, such as data, organization, address, etc, based on SpaCy;

3. **Chunking:** In order to avoid missing salient information units, chunking of noun phrases for entities and chunking of action related phrases for relations are performance in their work；

- **Noun chunks**: phrases that have a noun and the words describing the noun, a suburb of Detroit;
- **chunking of action words:** verb phrases can contain verbs, particles and/or adverbs that represent more meaningful relations between entities. For example, was founded by;

4. **Dependency parsing:** There exists semantic association (affiliation) between each words, using Spacy;

# Triple Extraction

Ford Motor Company is an American multinational automaker that has its main headquarters in Dearborn, Michigan, a suburb of Detroit. The company was founded by Henry Ford and incorporated on June 16, 1903.

| Sent # | Phrase # | Phrase | Type |
|---|---|---|---|
| 0 | 0 | Ford Motor Company | ENTITY |
| 0 | 1 | is | VERB |
| 0 | 2 | an American multinational automaker | ENTITY |
| 0 | 3 | that | DET |
| 0 | 4 | has | VERB |
| 0 | 5 | its main headquarters | ENTITY |
| 0 | 6 | in | ADP |
| 0 | 7 | Dearborn | ENTITY |
| 0 | 8 | , | PUNCT |
| 0 | 9 | Michigan | ENTITY |
| 0 | 10 | , | PUNCT |
| 0 | 11 | a suburb of Detroit | ENTITY |
| 0 | 12 | . | PUNCT |
| 1 | 13 | The company | ENTITY |
| 1 | 14 | was founded by | VERB |
| 1 | 15 | Henry Ford | ENTITY |
| 1 | 16 | and | CCONJ |
| 1 | 17 | incorporated on | VERB |
| 1 | 18 | June 16, 1903 | ENTITY |
| 1 | 19 | . | PUNCT |

**Algorithm 1** Chunking of noun phrases and verb phrases

**procedure** CHUNKPHRASES(*document*)
  **for each** *sentence in document* **do**
    ▷ Chunk noun phrases (NPs) and tag as *ENTITY*
      **chunk** *NPs* ▷ NP
      **chunk** $'('+NP+')'$ ▷ (NP)
      **chunk** $NP+' of'+NP$ ▷ NP of NP
      **chunk** $NP+NP$ ▷ NP NP
    ▷ Chunk verb phrases and tag as *VERB*
      **chunk** $VERB+PART$ ▷ verb + particle
      **chunk** $VERB+ADP$ ▷ verb + adpositions
      **chunk** $ADP+VERB$ ▷ adpositions + verb
      **chunk** $PART+VERB$ ▷ particle + verb
      **chunk** $VERB+VERB$ ▷ verb + verb
  **return** *document* ▷ Document with phrase chunks

# Triple Extraction

- ORG: Companies, agencies, institutions, etc
- NORP: Nationalities or religious or political groups
- GPE: Countries, cities, states;
- GPE: Geopolitical Entity
- NNP: proper noun
- VBZ: verb, 3rd person sing
- DT: determiner
- JJ: adjective
- NN: noun, singular
- WDT: wh-determiner which
- PRP:  personal pronoun
- IN: preposition/subordinating conjunction
- VBD: verb, past tense
- VBN: verb, past participle
- CC: coordinating conjunction
- CD: cardinal digit

| Token Id | Token | Entity Type | IOB | Coarse Grained POS | POS | Start | End | Dependency |
|---|---|---|---|---|---|---|---|---|
| 0 | Ford | ORG | B | PROPN | NNP | 0 | 3 | compound |
| 1 | Motor | ORG | I | PROPN | NNP | 5 | 9 | compound |
| 2 | Company | ORG | I | PROPN | NNP | 11 | 17 | nsubj |
| 3 | is | | O | VERB | VBZ | 19 | 20 | ROOT |
| 4 | an | | O | DET | DT | 22 | 23 | det |
| 5 | American | NORP | B | ADJ | JJ | 25 | 32 | amod |
| 6 | multinational | | O | ADJ | JJ | 34 | 46 | amod |
| 7 | automaker | | O | NOUN | NN | 48 | 56 | attr |
| 8 | that | | O | DET | WDT | 58 | 61 | nsubj |
| 9 | has | | O | VERB | VBZ | 63 | 65 | relcl |
| 10 | its | | O | DET | PRP | 67 | 69 | poss |
| 11 | main | | O | ADJ | JJ | 71 | 74 | amod |
| 12 | headquarters | | O | NOUN | NN | 76 | 87 | dobj |
| 13 | in | | O | ADP | IN | 89 | 90 | prep |
| 14 | Dearborn | GPE | B | PROPN | NNP | 92 | 99 | pobj |
| 15 | , | | O | PUNCT | , | 100 | 100 | punct |
| 16 | Michigan | GPE | B | PROPN | NNP | 102 | 109 | appos |
| 17 | , | | O | PUNCT | , | 110 | 110 | punct |
| 18 | a | | O | DET | DT | 112 | 112 | det |
| 19 | suburb | | O | NOUN | NN | 114 | 119 | dobj |
| 20 | of | | O | ADP | IN | 121 | 122 | prep |
| 21 | Detroit | GPE | B | PROPN | NNP | 124 | 130 | pobj |
| 22 | . | | O | PUNCT | . | 131 | 131 | punct |
| 23 | The | | O | DET | DT | 133 | 135 | det |
| 24 | company | | O | NOUN | NN | 137 | 143 | nsubjpass |
| 25 | was | | O | VERB | VBD | 145 | 147 | auxpass |
| 26 | founded | | O | VERB | VBN | 149 | 155 | ROOT |
| 27 | by | | O | ADP | IN | 157 | 158 | agent |
| 28 | Henry | PERSON | B | PROPN | NNP | 160 | 164 | compound |
| 29 | Ford | PERSON | I | PROPN | NNP | 166 | 169 | pobj |
| 30 | and | | O | CCONJ | CC | 171 | 173 | cc |
| 31 | incorporated | | O | VERB | VBD | 175 | 186 | conj |
| 32 | on | | O | ADP | IN | 188 | 189 | prep |
| 33 | June | DATE | B | PROPN | NNP | 191 | 194 | pobj |
| 34 | 16 | DATE | I | NUM | CD | 196 | 197 | nummod |
| 35 | , | DATE | I | PUNCT | , | 198 | 198 | punct |
| 36 | 1903 | DATE | I | NUM | CD | 200 | 203 | nummod |
| 37 | . | | O | PUNCT | . | 204 | 204 | punct |

https://spacy.io/

# Triple Extraction

5. **Coreference Resolution :** A list of coreferenced items is created using NeuralCoref. Ford Motor Company - The company. In the case of pronouns such as its, her, his or their, we ignore the coreference items;
6. **Triple Mapping :** extract triples;

**Algorithm 2** Triple mapping algorithm

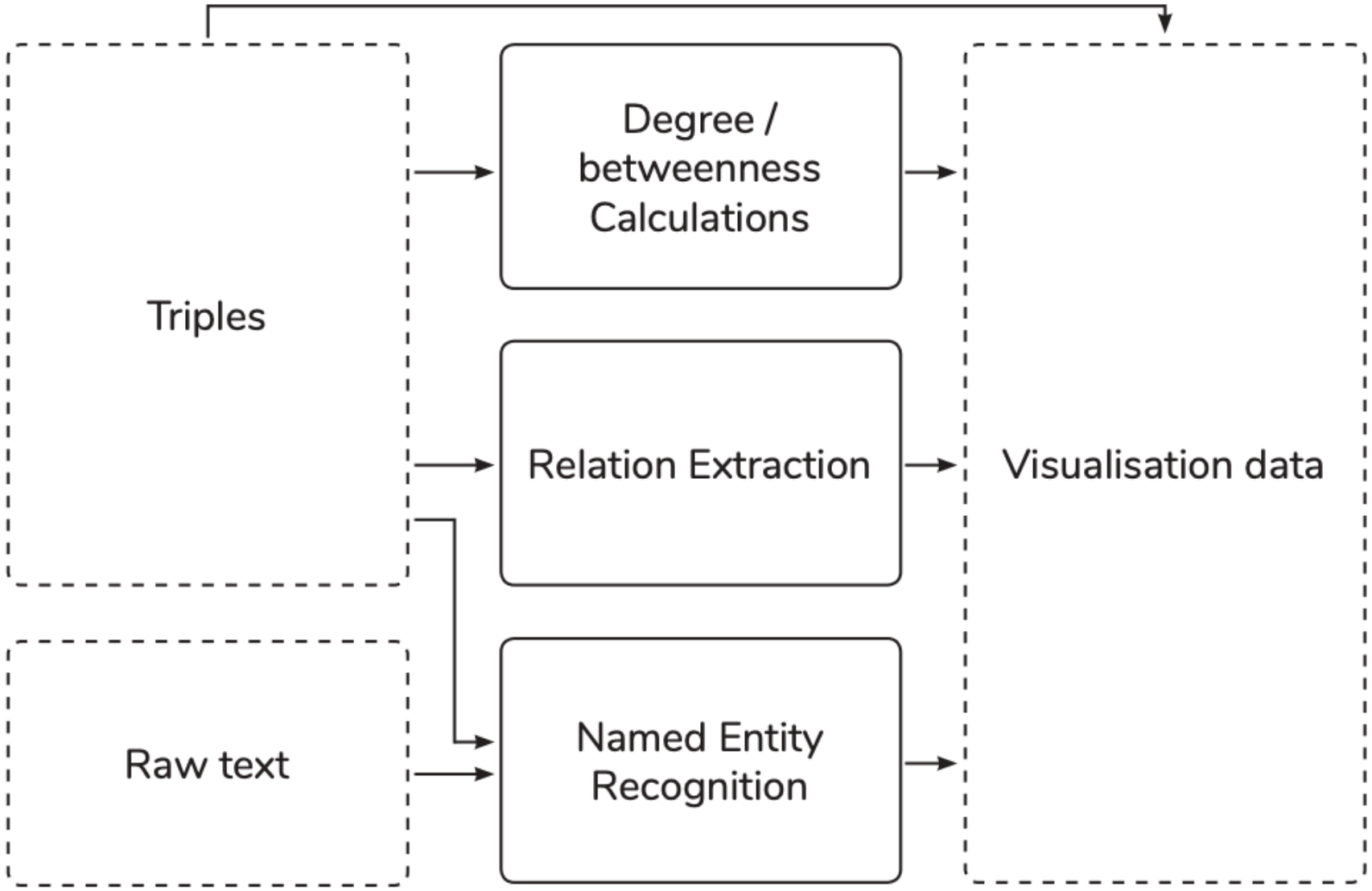
```
procedure GETTRIPLES(document)
    for each sentence in document do
        relations ← verbs + prepositions + postpositions      ▷ Select relations such as showcased, has, in, to, during
        for each r in relations do
            heads ← entities on the left side of r            ▷ Get the head entities for the relation r
            tails ← entities on the right side of r           ▷ Get the tail entities for the relation r
            for each h in heads do
                for each t in tails do
                    triples ← triples + [h, r, t]             ▷ Add [head, relation, tail] to the list of triples
    return triples                                            ▷ Return the list of triples
procedure EXTRACTTRIPLES(document)
    ▷ Extract triples from the document at the sentence level
    triples ← GETTRIPLES(document)
    ▷ Extract the triples at the document level using the graph shortest paths
    G ← create graph(triples)                                 ▷ Build a graph from the triples using NetworkX package
    paths ← get shortest paths(G)                             ▷ Get all shortest paths between named entities
    for each h, t in pairs of named entities do
        if h and t connected by a path using 'in', 'at', 'on' prepositions then
            triples ← triples + [h,' in', t]                  ▷ Add [head, 'in', tail] to the list of triples
    return triples                                            ▷ Return the full list of triples
```

# Triple Extraction

7. **Triple Filtering:** To improve the quality of the triples, the filtering is performed to remove any triple with a stop word as a head entity. The stop words include NLTK stop words, names of days (Monday to Sunday) and names of months (January to December);
8. **Article Removal:** To clean the entities we removed some tokens including articles (e.g., a, an, the), possessive pronouns (e.g., its, their) and demonstrative pronouns (e.g., that, these) from the head and tails of each triple;

# Visualization system

- **Degree/betweenness calculation:** determines the degree and betweenness centrality of the head and tail of each triple;
- **Relation extraction:** component maps the relation phrase of each triple to one or more structured relation types;
- **Named entity recognition:** component determines the semantic type of the head and tail of each triple;



# Visualization system

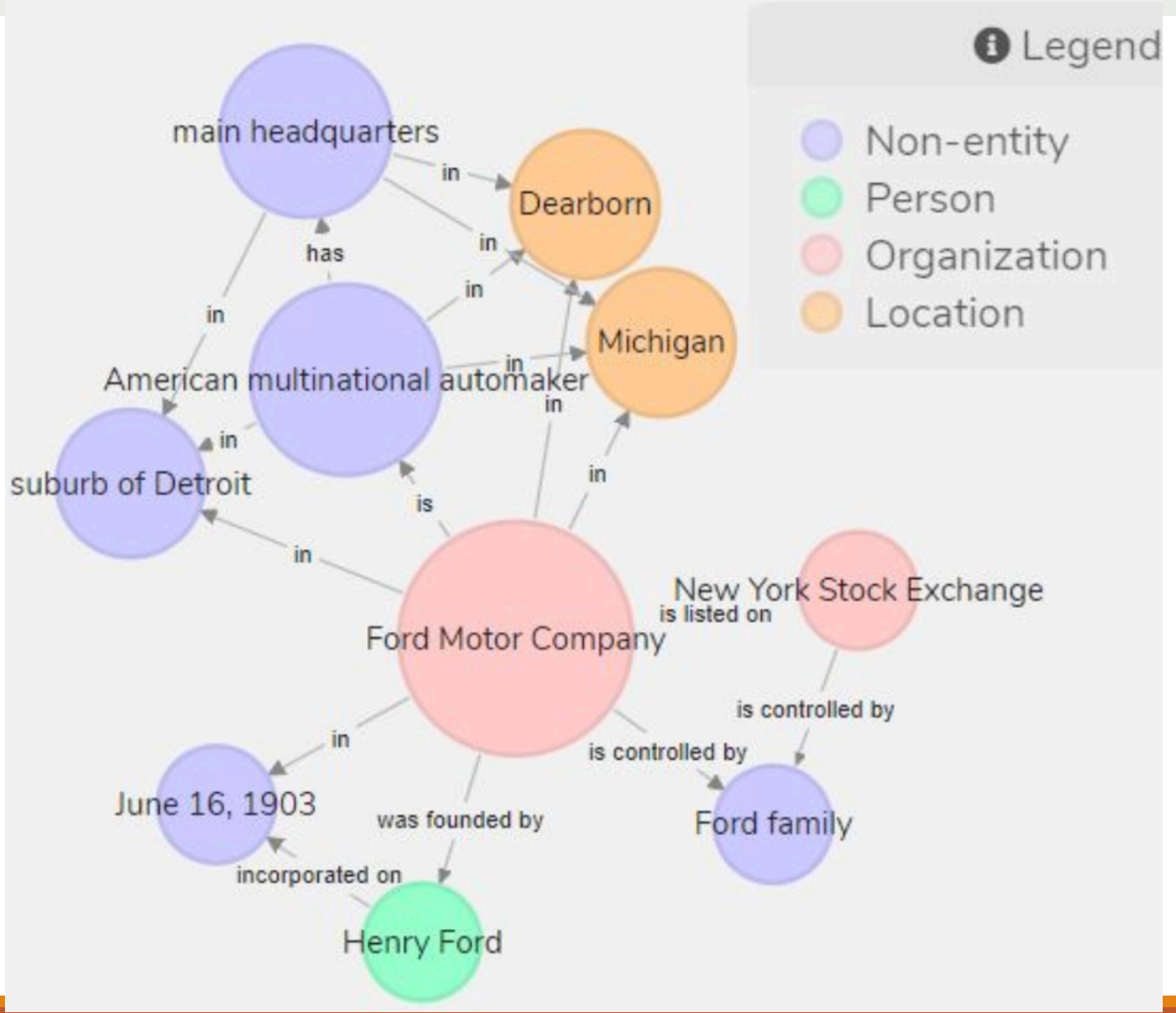
1. **Degree/betweenness calculation:** Degree refers to the number of edges connected to a node. Betweenness calculation measures the extent to which each vertex lies along the paths between other vertices, both of which reflects the importance of entity;
2. **Relation extraction:** use an Att-LSTM model which maps a sequence of words padded with entity (the head and tail of each triple) markers to a fixed relation type. To be specific feed them into a pretrained model (trained on the SemEval 2010 Task 8 dataset) to obtain the corresponding SemEval relation;
3. **Named entity recognition:** mapping SpaCy tagging to Wikipedia NER (PER, ORG, LOC, MISC, and O) by edit distance to allow for a greater level of abstraction and flexibility;

| Triple | | | Additional information | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Head ($H$) | Relation ($R$) | Tail ($T$) | SemEval Relation | $Type_H$ | $Type_T$ | $Deg_H$ | $Deg_T$ | $Betw_H$ | $Betw_T$ |
| Ford Motor Company | in | Dearborn | Content-Container | ORG | LOC | 6 | 3 | 11.0 | 0.75 |
| Ford Motor Company | in | Michigan | Content-Container | ORG | LOC | 6 | 3 | 11.0 | 0.75 |
| Ford Motor Company | in | suburb of Detroit | Member-Collection | ORG | O | 6 | 3 | 11.0 | 0.75 |
| Ford Motor Company | in | June 16, 1903 | Component-Whole | ORG | O | 6 | 2 | 11.0 | 0.0 |
| Ford Motor Company | is | American multinational automaker | Instrument-Agency | ORG | O | 6 | 5 | 11.0 | 1.75 |
| Ford Motor Company | was founded by | Henry Ford | Product-Producer | ORG | PER | 6 | 2 | 11.0 | 0.0 |
| American multinational automaker | in | Dearborn | Member-Collection | O | LOC | 5 | 3 | 1.75 | 0.75 |
| American multinational automaker | in | Michigan | Member-Collection | O | LOC | 5 | 3 | 1.75 | 0.75 |
| American multinational automaker | in | suburb of Detroit | Member-Collection | O | O | 5 | 3 | 1.75 | 0.75 |
| American multinational automaker | has | main headquarters | Cause-Effect | O | O | 5 | 4 | 1.75 | 1.0 |
| Henry Ford | incorporated on | June 16, 1903 | Component-Whole | PER | O | 2 | 2 | 0.0 | 0.0 |
| main headquarters | in | Dearborn | Content-Container | O | LOC | 4 | 3 | 1.0 | 0.75 |
| main headquarters | in | Michigan | Content-Container | O | LOC | 4 | 3 | 1.0 | 0.75 |
| main headquarters | in | suburb of Detroit | Member-Collection | O | O | 4 | 3 | 1.0 | 0.75 |

# Visualization system

Ford Motor Company is an American multinational automaker that has its main headquarters in Dearborn, Michigan, a suburb of Detroit. It was founded by Henry Ford and incorporated on June 16, 1903.

- The nodes are coloured based on their named entity types;
- The node sizes are based on their degree centrality values;




https://github.com/Michael-Stewart-Webdev/text2kg-visualisation

# Conclusion

- The system uses a pipeline-based approach to extract a set of triples from a given document. It offers a simple and effective solution to the challenge of knowledge graph construction from domain-specific text;

- It also provides the facility to visualize useful information about each triple such as the degree, betweenness, structured relation type(s), and named entity types;

- As for Chinese corpus, it has some modify to the pipeline, while the process may remain consistent;

# Reference

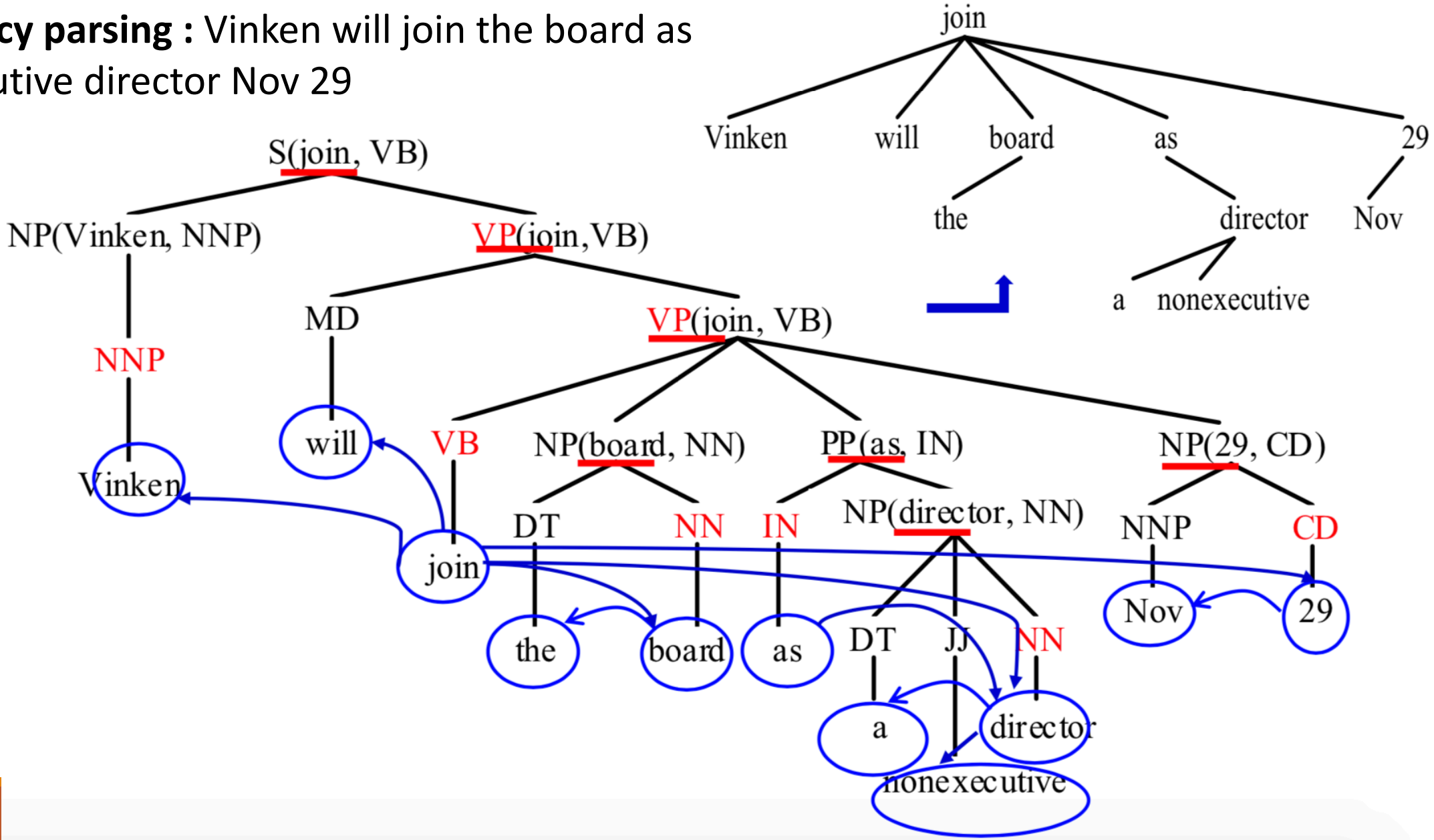

- Stewart M, Enkhsaikhan M, Liu W. ICDM 2019 Knowledge Graph Contest: Team UWA[J]. arXiv preprint arXiv:1909.01807, 2019.
- Triple extraction system: NLTK (https://www.nltk.org/) and SpaCy (https://spacy.io/).
- Visualization system is written in Flask7 (https://www.fullstackpython.com/flask.html).
- The front-end visualizations are written primarily in D3.js8 (https://d3js.org/).
- The attentionbased Bi-LSTM for relation extraction is implemented in Tensorflow, and trained on the SemEval 2010 Task 8 dataset.
- P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, “Attentionbased bidirectional long short-term memory networks for relation classification,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 207–212.
- The degree and betweenness calculations are performed via NetworkX9 (https://networkx.github.io).

# Q&A

## Thanks

# Appendiex

**Dependency parsing :** Vinken will join the board as a nonexecutive director Nov 29



# Appendiex

**Attention-based LSTM model**

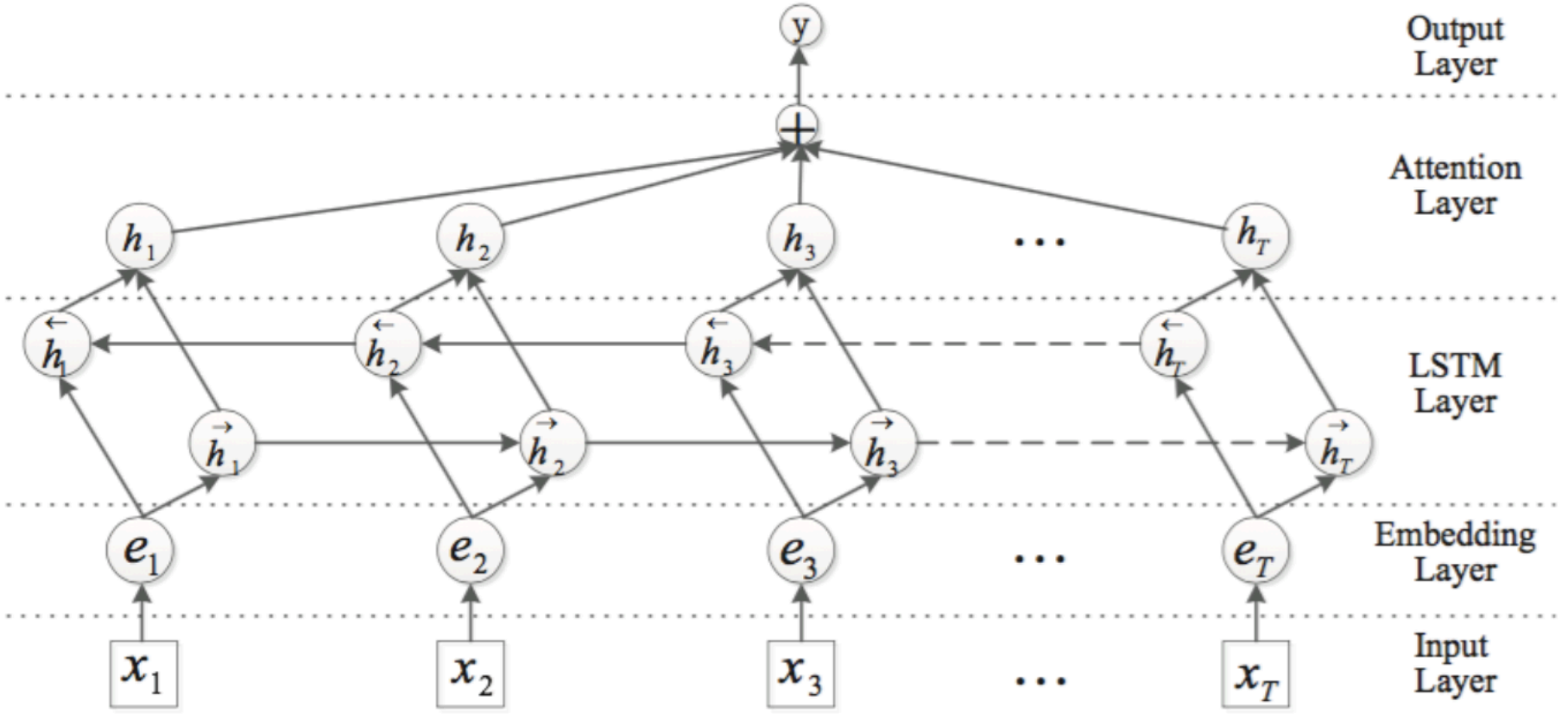



Figure 1: Bidirectional LSTM model with Attention