ICDM 2019

2019 ICDM/ICBK Knowledge Graph Contest

Xindong Wu

Mininglamp Academy of Sciences

ICDM 2019

Automatic Knowledge Graph Construction

- Build a knowledge graph from unstructured text in a specific domain or cross multiple domains, without human intervention.

ICDM 2019

Specification

Teams from both degree-granting institutions and industrial labs are invited;

A regular competition with a demonstration requirement for candidates willing to compete for a prize;

Supported by Mininglamp Academy of Sciences and Hefei University of Technology.

ICDM 2019

Background

- **Predominant studies**: developing a pipeline of NLP tasks
  - named entity recognition (NER);
  - co-reference resolution (CoRef),;
  - entity linking (EL); and
  - relation extraction (RE).

- **Few pay attention to how human knowledge is structured**
  - It is difficult to accurately reflect, in advance, the relations that will appear in someone's mind when they read a piece of text.

Data Graph vs. Knowledge Graph

- 99% of existing "knowledge graphs" are actually **data graphs** without knowledge.
- **Example**: "Bob and I were high-school classmates, and I will invite him for a dinner to celebrate our 25th year class reunion in 2020"
- If a graph cannot identify who is "him" and does not provide any support on when they graduated from high school, the graph is only a data graph.

Definition of a knowledge graph

- A knowledge graph as a semantic graph for describing concepts and their relations in a physical world with three essential components:
  - **Concepts**. A concept can be an entity (such as a person), an attribute (such as age), or a fact (such as "a red car with 4 doors"), and is represented as a node.
  - **Relations**. A relation is a connection between two nodes with a semantic label, such as "is-a", "has-a" or action (*e.g.*, "becomes").
  - **Background knowledge about concepts and relations**.
    - A concept can have
      - different names such as Professor X. Wu and Dr. Xindong Wu,
      - and possibly multiple attributes such as height and occupation,
    - A relation can have different appearances such as "had", "has" and "have".
    - The background knowledge in the form of a dictionary or an ontology can semantically link different names, attributes and appearances.

ICDM 2019

Challenges - *Knowledge Graph Construction*

- **Information loss:**
  - when the output graph is incomplete.
- **Information redundancy:**
  - extra concepts and relations that do not exist in the input text but in the background knowledge.
  - *example: "Bob hit the nail into the wall with a hammer."*
  - *Bob*, *nail*, *wall* and *hammer*, and the following relations: *(Bob, hit, nail)*, *(nail, into, wall)*, and *(Bob, with, hammer)*.
- **Information overlapping:**
  - whether a knowledge graph can encode the changing of an attribute.
  - *example: "John had a new fast 4-wheel car, and the car became a slow one 2 years later."*
  - *Bob*, *nail*, *wall* and *hammer*, and the following relations: *(Bob, hit, nail)*, *(nail, into, wall)*, and *(Bob, with, hammer)*.

ICDM 2019

Data

The dataset consists of 300 recent published articles from news media of 4 industries:

automotive engineering	cosmetics	public security	catering services

Each article is of 150 to 250 words, contains around 8-20 entities.

ICDM 2019

Rules

- Open to the public, with the following restrictions:
  - Organizers, contest committee members, and employees of Mininglamp **are not eligible** to participate in the contest.
  - Privately sharing code or data outside of teams **is not permitted**. It is okay to share code if made available at a public website.
  - Each participant can join **only one team**. The maximum number of members on each team is five.

Schedule and readiness

10 June 2019

Contest requirement specification and sample data available

20 June 2019

Team registration begins. Data sets are available

31 July 2019, 23:59

Final submission deadline: 23:59, July 31, 2019 (Pacific Time)

10 Aug. 2019, 23:59

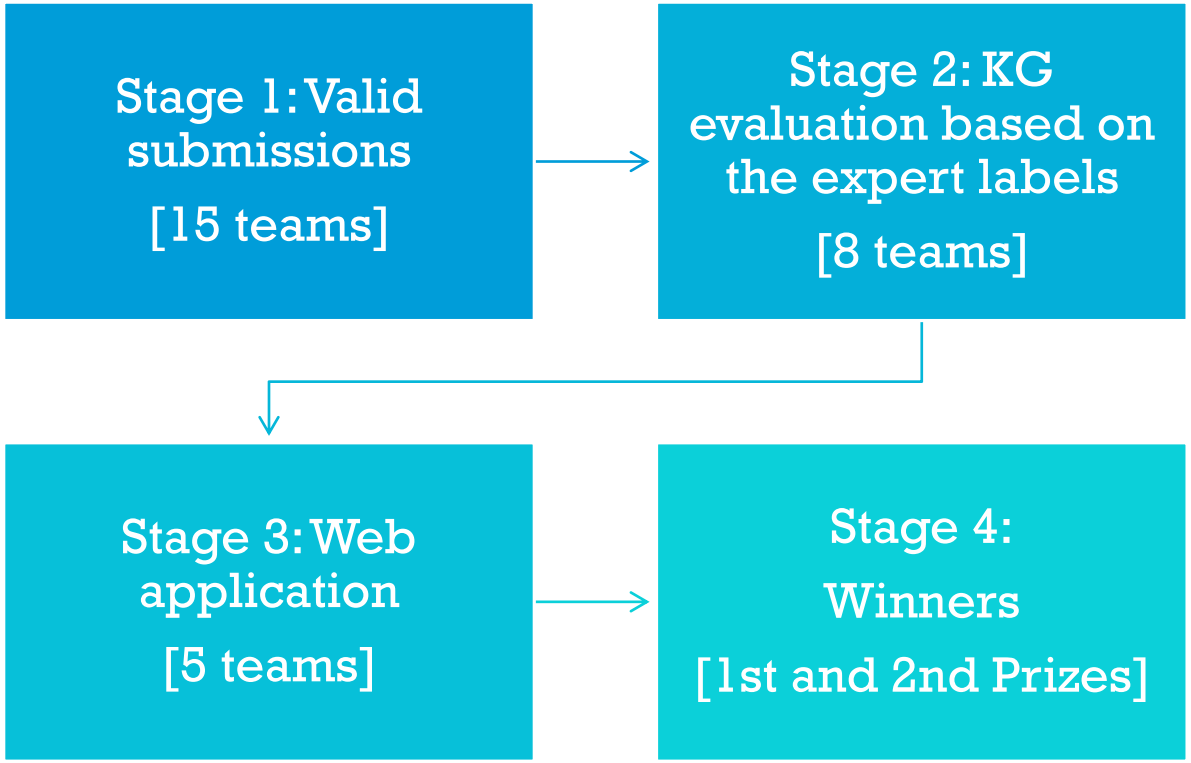
Notifications to shortlisted teams for web applications: 23:59, August 10, 2019 (Pacific Time)

30 Aug. 2019

Finalist notifications

ICDM 2019

Stages



Prizes

TOP 5 TEAMS OF 2019 ICDM/ICBK CONTEST

Team	Award	Organization
UWA	First Prize	University of Western Australia, Australia
Tmail	Second Prize	Huazhong University of Science and Technology, and Tencent Medical AI Lab, China
BUPT-IBL	Honorable Mention	Beijing University of Posts and Telecommunications, China
MIDAS-IIITD	Honorable Mention	Indraprastha Institute of Information Technology Delhi, India
Lab1105	Honorable Mention	Wuhan University of Technology, China

- First Prize: USD 10,000
- Second Prize: USD 3,000
- Honorable Mention: One free Registration (equivalent to USD 962)

ICDM 2019

Key Techniques

A typical knowledge-graph construction process consists of three main components:

information extraction	knowledge fusion	knowledge processing

This Contest only involves information extraction and knowledge fusion.

ICDM 2019

Entity Recognition

- *Rule-based approaches*
  - manually construct a limited set of rules and then search for strings in the text that matched these rules.
- *Machine learning (ML)-based approaches*
  - the selection of ML models and methods, the improvement of models and methods, and the selection of features
- *Deep learning-based approaches*

ICDM 2019

Main Relation Extraction Methods -Winning Teams

Team UWA	NLP tool SpaCy chunked noun and verb phrases extracted according to predefined rules.
Team Tmail	Stanford OpenIE toolkit, OpenIE 5.0, and SpaCy.
Team BUPT-IBL	Their own model SC-LSTM, Stanford CoreNLP, and SpaCy.
Team MIDAS-IIITD	NLTK and SpaCy, NLP toolkit flair.
Team Lab1105	SpaCy, a BiLSTM+CRF model.

Relation Extraction

- *Supervised learning methods:*
  - recognize entities through a matching process and extract specific relations.
- *Semi-supervised methods:*
  - alleviate dependence on huge amounts of labels.
- *Domain-independent methods:*
  - relax the need for domain specifications.
- *Distant-supervised methods*
  - generate large amounts of training data automatically by aligning unstructured text with a knowledge base.
- *Deep learning methods*

ICDM 2019

Main Relation Extraction Methods -Winning Teams

Team UWA	A pre-trained attention-based Bi-LSTM model.
Team Tmail	Stanford OpenIE toolkit, OpenIE 5.0, and SpaCy.
Team BUPT-IBL	Stanford OpenIE toolkit, a model based on a syntax-tree.
Team MIDAS-IIITD	Hand-written rules.
Team Lab1105	SpaCy and hand-written rules.

ICDM 2019

Co-reference Resolution

- Co-reference resolution or entity resolution is used when an entity in a knowledge base is linked to multiple entity references.
- For instance, "President Trump" and "Donald John Trump" are the same person, so these two entity references should be merged before they are connected to an entity in the knowledge base.
- Most solutions to entity resolution have been based on state-of-the-art machine learning methods

Main Co-reference Resolution Methods -Winning Teams

- Four of the five winning teams in this Contest (UWA, BUPT-IBL, MIDAS-IIITD, and Lab1105) used NeuralCoref [1] for entity resolution.

- [1] T. Wolf, “Neuralcoref 4.0: Coreference resolution in spacy with neural networks.” 2017.

Acknowledgements

- We would like to thank Mininglamp Technology (Beijing, China) for preparing labeled datasets and the contest website, the Contest committee for the evaluations of submissions, and all teams for their participation.

- More details can be found in :
  - X. Wu *et al.*, Automatic Knowledge Graph Construction: A Report on the 2019 ICDM/ICBK Contest, ICDM 2019, pp. 1540-1545.