文章阅读方法

Technical Paper Guidelines

- What is the ***research aim*** of this paper? What ***problem*** is the approach trying to solve?
- Which ***method*** proposed in this paper to solve the problem? What are the main equations and/or algorithms in the paper?
- What is the ***information or external source*** used in their method?
- What is the ***main innovation*** of the paper? How does it *relate to previous work*?
- Is the author's ***description*** of the previous work accurate? or misleading?
- What are the ***conclusions*** of this paper?

文章阅读方法（续）

Data Analysis Guidelines

* Why is this task difficult?
* What are the hard cases?
* What are the easy cases?
* Do these models make ***assumptions*** to solve the task?
* What are the ***conclusions*** of data analysis?

实例介绍

Title: Result of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition

* Abstract: Problems, methods, results

文章解决什么问题？

如何解决的？

文章的结论或结果是什么？

实例介绍（续）

* Aim:
1. Provide **a definition** of **emerging** and **rare** entities, and **datasets** for detecting these entities.
2. Evaluate the **ability** of participating entries to detect and classify novel and emerging named entities in noisy text.
* Problem: recall on **unusua**l, previous-**unseen** entities in the **noise** text is low.
* Information collect:
1. three different sources, Reddit, Twitter, YouTube as dev. and test data.
2. StackExchange (including movies, politics, physics, sci-fi and security)
* Methods: to develop **systems** that are less sensitive to change, and can handle rare and emerging entity types with ease.

实例介绍（续）

* Reason & assumption（我的理由和假设是什么）:

to develop systems that are less sensitive to change, and can handle rare and emerging entity types with ease.

* Analysis results:

1. F1(entity,type) surface recognition is lower than entity recognition.
2. NER in novel emerging settings remains **hard**:

- When a name is a common word(Donald Duck), some system only identify part of the name
- Locations that contain elements that are also common in person names present an obstacle for the participating systems
- Names originating from other languages often present problems for the systems.
- Corporation and creative work were generally a difficult classes for the system to predict.

3. Annotation remains hard.

实例介绍（续）

* Conclusion:

1. **A new benchmark dataset** consisting of 1008 development and 1287 test documents containing nearly 2000 entity mentions.
2. 7 systems show that entity recognition on these entities is **more difficult** than high frequent entities commonly found in NER.
3. systems are failing to generalize successfully, instead profiting from frequently repeated entities in regular contexts.