Electrical & Computer ENGINEERING

A Neural Network-Powered Cognitive Method of Identifying Semantic Entities in Earth Science Papers

Xiaoyi Duan, Jia Zhang
Carnegie Mellon University
Mountain View, CA 94087
{xiaoyi.duan;jia.zhang}@sv.cmu.edu

Rahul Ramachandran, Patrick Gatlin, Manil Maskey
NASA/MSFC
Huntsville, AL 35811
{rahul.ramachandran;patrick.gatlin;manil.maskey}@nasa.gov

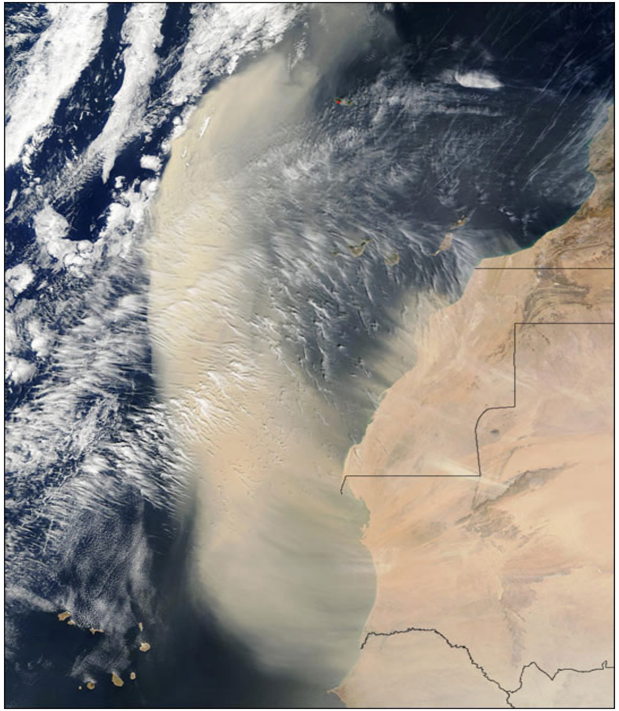
Jeffrey J. Miller, Kaylin Bugbee
University of Alabama in Huntsville
Huntsville, AL 35811
{jjm0022;kmb0100}@uah.edu

Tsengdar J. Lee
Science Mission Directorate, NASA Headquarters
Washington, D.C. 20546
tsengdar.j.lee@nasa.gov

Carnegie Mellon University

Outline

- **Introduction**
  - **Motivation**
  - **Related work**
- Our work
  - Profile-matching
  - Neural Network
- Experiments
- Conclusion



Motivation

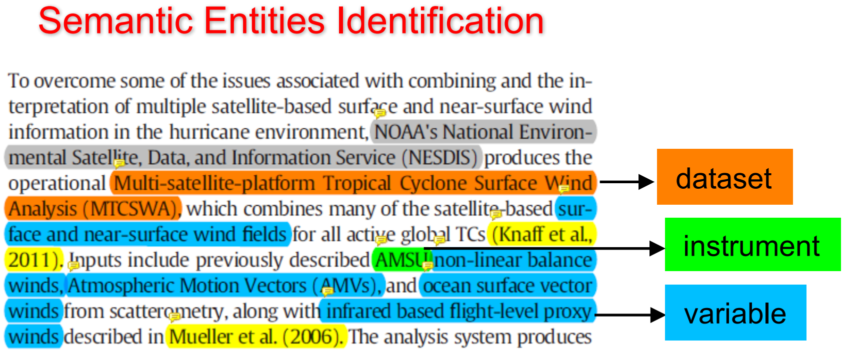
- ❑ Knowledge Explosion of academic publications
  - ❑ Learn from how human read earth science papers.
  - ❑ Machine helps human identify useful information.


Remote Sensing of Environment

Remote Sensing of Environment

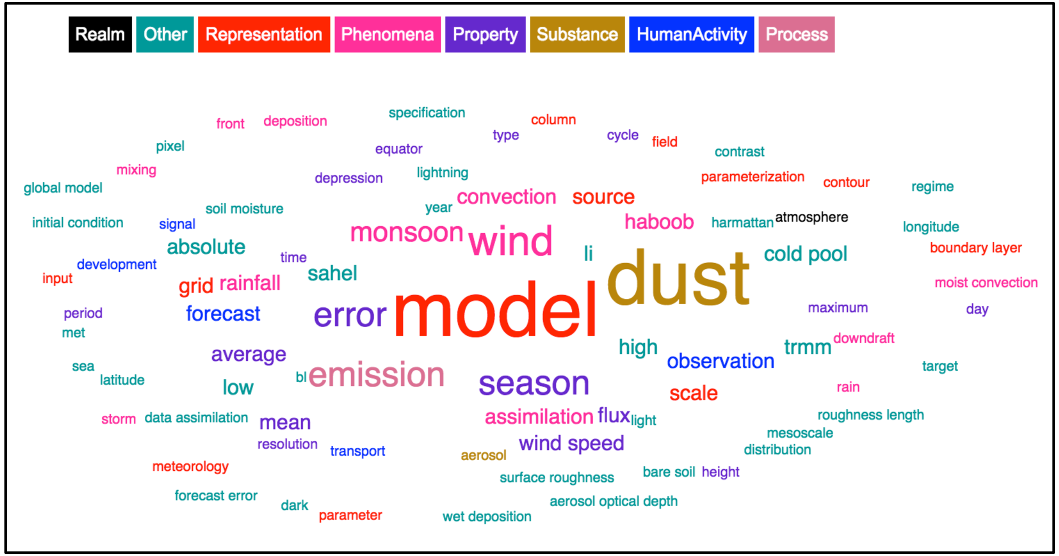
Improved algorithm for MODIS satellite retrievals of aerosol optical thickness over land in dusty atmosphere: Implications for air quality monitoring in China

Jun Wang, Xiaoguang Xu, Robert Spurr, Yuxuang Wang, Easan Drury



Motivation

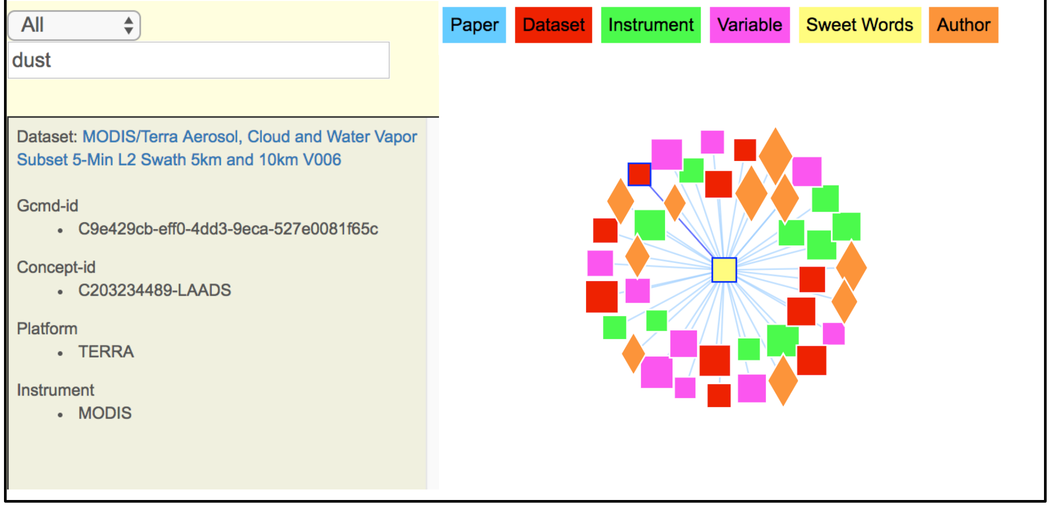
- ❑ Applications based on Semantic Entity Identification
  - ❑ Word cloud
  - ❑ Knowledge network
  - ❑ QA system
  - ❑ …



Motivation

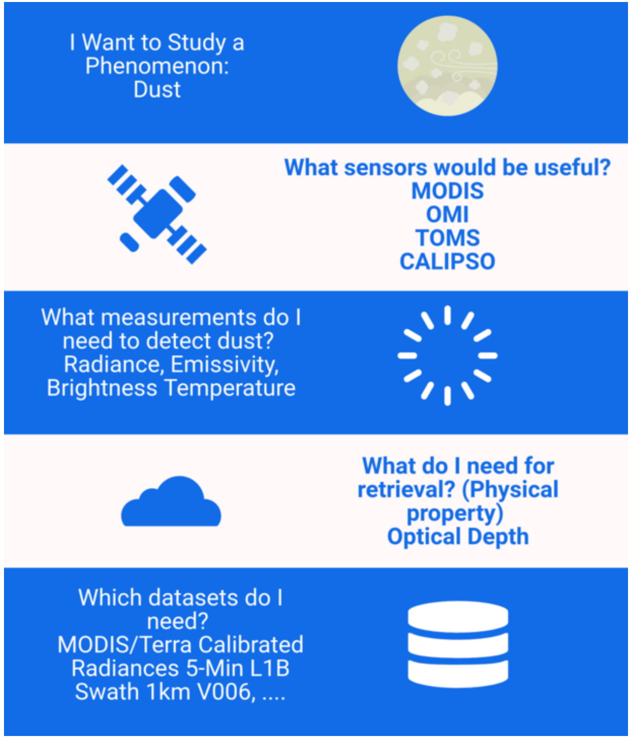
- ❑ Applications based on Semantic Entity Identification
  - ❑ Word cloud
  - ❑ Knowledge network
  - ❑ QA system
  - ❑ …

Query: What entities are most studied for topic 'dust'?



Motivation

- Applications based on Semantic Entity Identification
  - Word cloud
  - Knowledge network
  - QA system
  - …



Challenges

- ❑ Unstructured information
  - ❑ Text, table, caption, …
- ❑ Many ways to describe a thing
  - ❑ Dataset is uniquely identified by DOI (Digital Object Identifier)
- ❑ Unlikely to manually label data
  - ❑ Big volume
  - ❑ Domain knowledge

To overcome some of the issues associated with combining and the interpretation of multiple satellite-based surface and near-surface wind information in the hurricane environment, NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) produces the operational Multi-satellite-platform Tropical Cyclone Surface Wind Analysis (MTCSWA), which combines many of the satellite-based surface and near-surface wind fields for all active global TCs (Knaff et al., 2011). Inputs include previously described AMSU non-linear balance winds, Atmospheric Motion Vectors (AMVs), and ocean surface vector winds from scatterometry, along with infrared based flight-level proxy winds described in Mueller et al. (2006). The analysis system produces

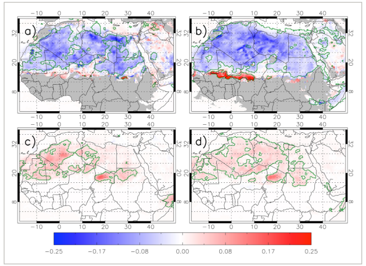
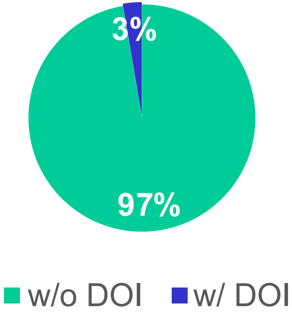


Figure 2 Open in figure viewer | PowerPoint

The influence of wind regimes. Model DAI AOD composited under (a and b) high (>7 m/s) and (c and d) low (<7 m/s) model 10 m wind speeds minus the average DAI (see Figures 1c and 1d) during the monsoon season (Figures 2a and 2c) and the nonmonsoon season (Figures 2b and 2d). Green contouring shows significant (95%) differences from the seasonal average DAI using a bootstrapping method (see supporting information).

NASA SEDAC[1] dataset citations



1. http://sedac.ciesin.columbia.edu/

Related Work in Knowledge Extraction

- Existing Knowledge Base System
  - Google Knowledge Graph
  - Deep Dive
  - Microsoft Academic Graph
  - IBM Watson
- Semantic entity extraction methods
  - Named entity recognition
  - Unsupervised learning

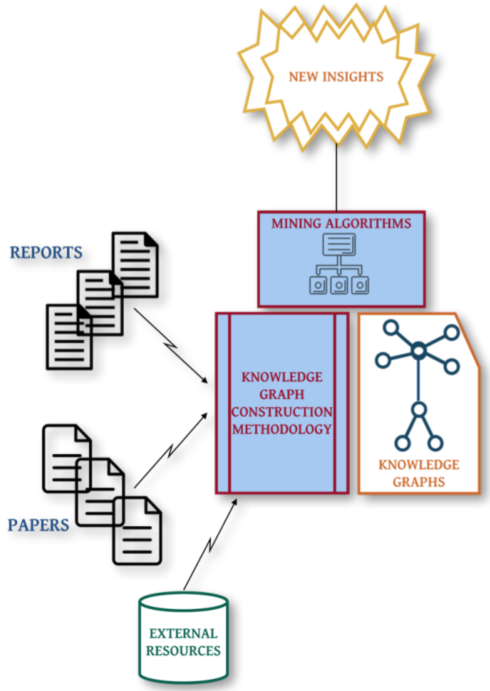
Extensive human involvement

Demand a lot of labeled data

Lack of accuracy

Outline

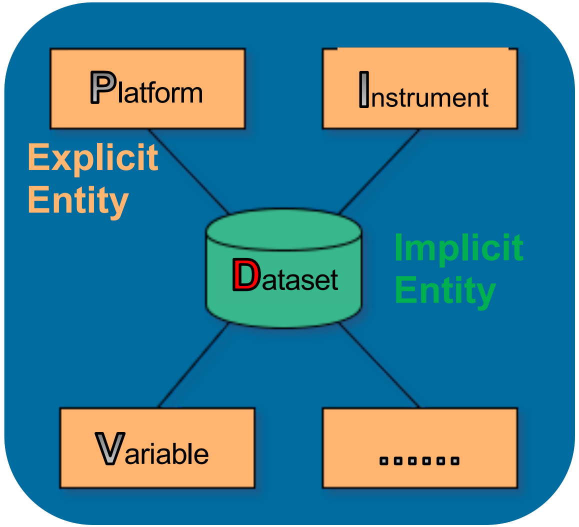
- ❑Introduction
  - ❑Motivation
  - ❑Related work
- **❑Our work**
  - **❑Profile-matching**
  - **❑Neural Network**
- ❑Experiments
- ❑Conclusion



Problem Definition

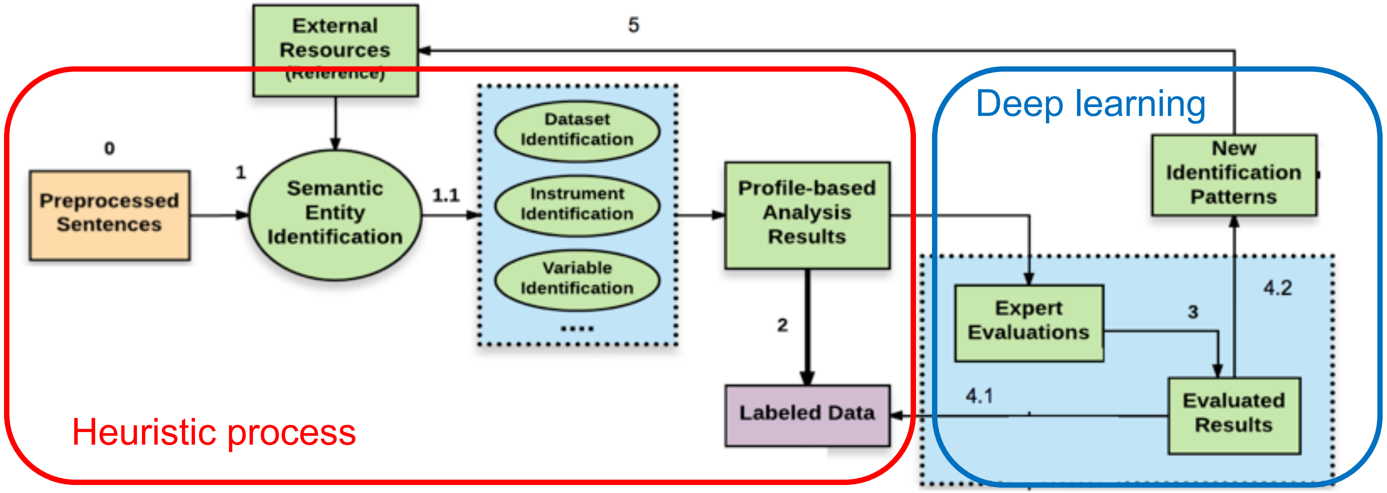
- ❑ **Explicit Entity** is cited by certain names.
- ❑ **Implicit Entity** is usually mentioned implicitly and described by sentences in close proximity to the entity.

- ❑ **Semantic Entity Identification for Earth Science**: automatically identify key semantic entities from contents of paper, where explicit entities are from I, V or P, and implicit entities are from D.



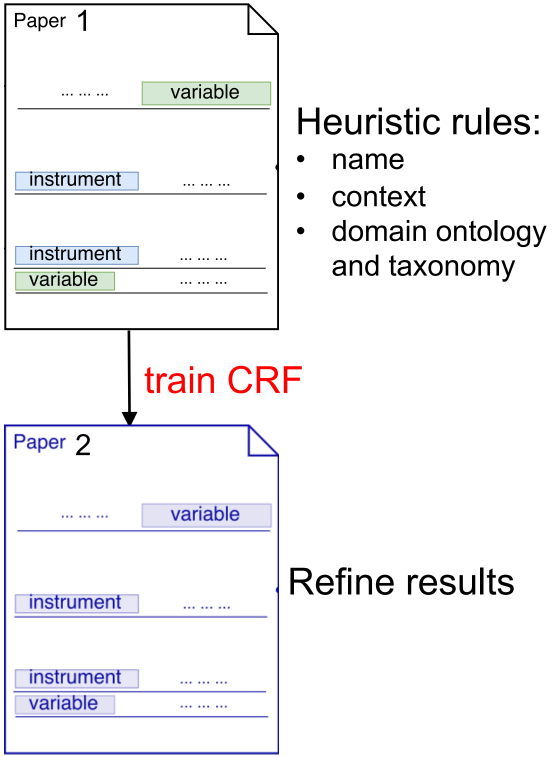
Framework of Semantic Entity Identification

1. Heuristic algorithms for semantic entity identification to build a large training set [Steps 0-2]
2. Deep learning algorithms to improve results [Steps 3-5]



Heuristic-based Extraction

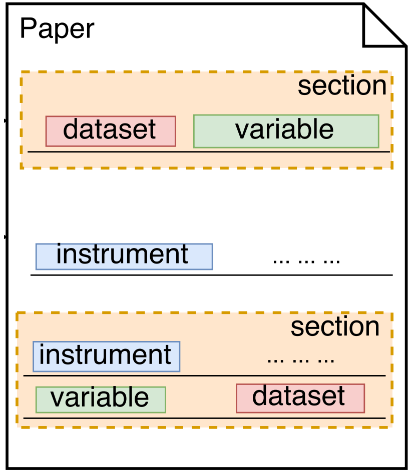
- ❑ Extract explicit entities
  - ❑ Heuristic rules
  - ❑ Train Conditional Random Field (CRF)[2] model
- ❑ Instrument and Platform: S(L), L(S)
- ❑ Variable v: {topic→term}1..*
  - *E.g. "rainfall amount"*: {Precipitation→Precipitation Amount}, {Precipitation→Rain}



2. J. R. Finkel, T. Grenager, and C. D. Manning, Incorporating non-local information into information extraction systems by gibbs sampling," ACL 2005, pp. 363–370.

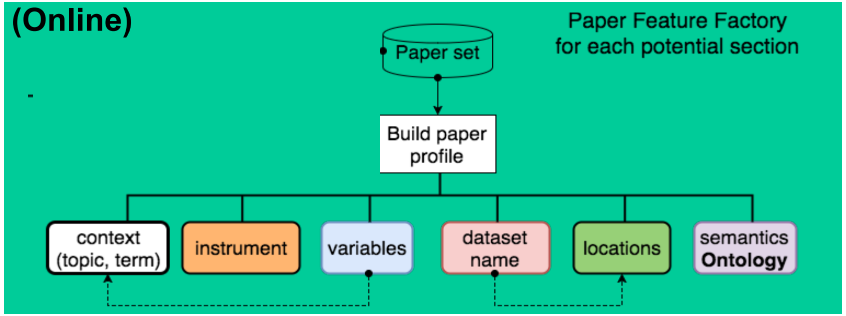
Weighted-Profile-Matching Dataset Extraction

- Observation: datasets are typically mentioned surrounded by some explicit entities
- Dataset identification
  - Extract potential section which may contain dataset
  - Compare section with every dataset and select the most similar one

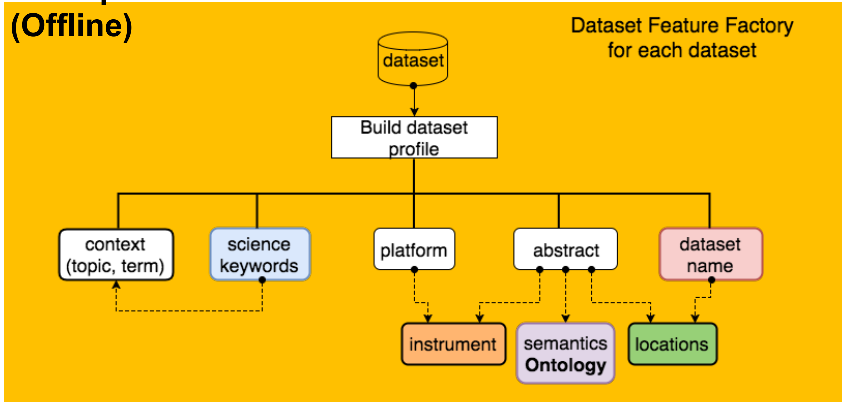


Weighted-Profile-Matching Dataset Extraction

Build profile for paper section (Online)



Build profile for dataset (Offline)



- Find the most relevant dataset with the highest weighted score
  - w_e: weight of entity *e*, and $w_i + w_v + w_p = 1$
  - S_{ed}: similarity between entity *e* in the section and dataset profile *d*, normalized to [0,1].

$$S_d = w_i \cdot S_{id} + w_v \cdot S_{vd} + w_p \cdot S_{pd}$$

- Hard to predefine the attribute weights

Neural Network-Powered Entity Extraction

- ❑ continuous bag-of-words (CBOW) model
  - ❑ Word embedding in NLP

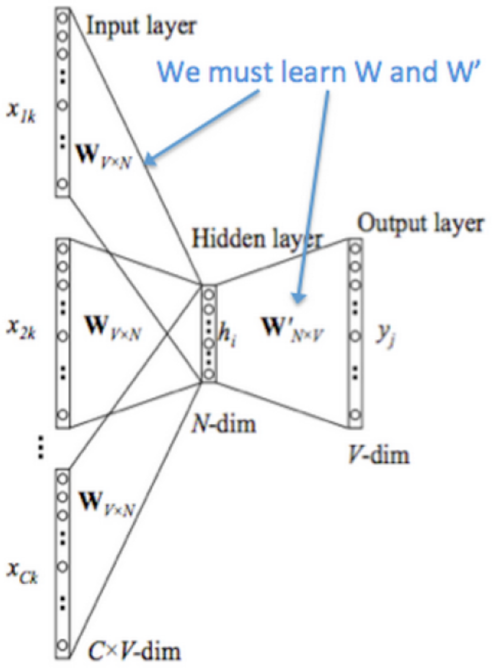
Input text corpus

- Computational lens on big social and information networks.
- The connections between individuals form the structural …
- In a network sense, individuals matters in the ways in which …
- Accordingly, this thesis develops computational models to investigating the ways that …
- We study two fundamental and interconnected directions: user demographics and network diversity
- … …

Train every context window

WHAT	IF	MIKE	WAS	SHORT	FOR	MICYCLE
WHAT	IF	MIKE	WAS	SHORT	FOR	MICYCLE
WHAT	IF	MIKE	WAS	SHORT	FOR	MICYCLE
		W_{c-2}	W_{c-1}	W_c	W_{c+1}	W_{c+2}

Learn word embedding W



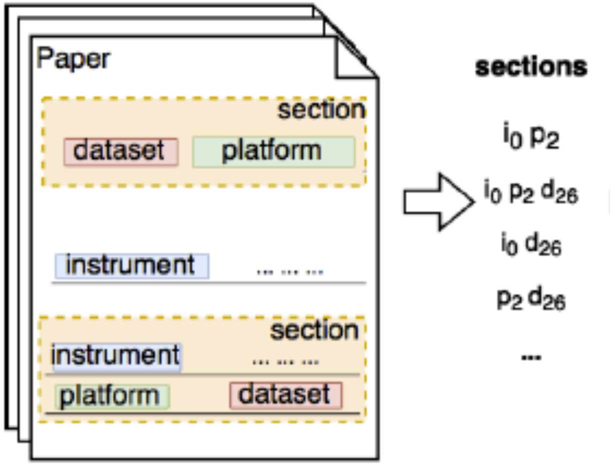
Optimization Objective:

$$\begin{aligned}\text{minimize } J &= -\log P(w_c|w_{c-m},\ldots,w_{c-1},w_{c+1},\ldots,w_{c+m})\\ &= -\log P(u_c|\hat{v})\\ &= -\log \frac{\exp(u_c^T\hat{v})}{\sum_{j=1}^{|V|}\exp(u_j^T\hat{v})}\end{aligned}$$

Neural Network-Powered Entity Extraction

- Neural Network Entity Extraction (NNEE): applies the CBOW model to predict a dataset from the surrounding (geographically close) context of entities.

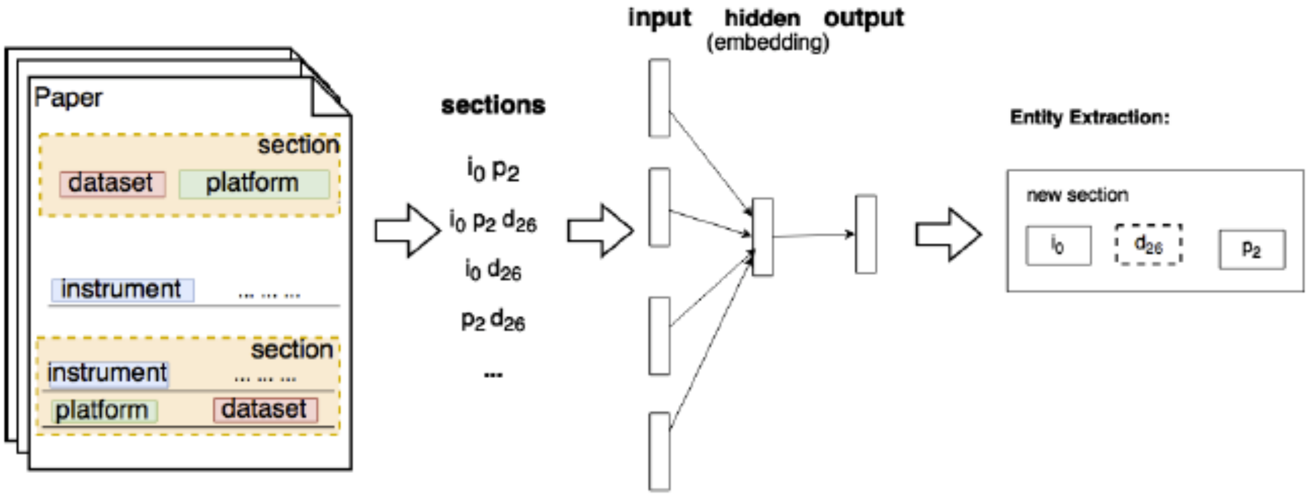
- Every entity is regarded as a word.
- Explicit entities in one identified potential area make up one sentence to train the NNEE model.



Neural Network-Powered Entity Extraction

- Cost function: Maximize the log probability of the dataset given any context entities.

$$J = -\log p(d|c_d) \qquad p(d|c_d) = \sigma(E'^T_d \cdot c_d) = \frac{exp(E'^T_d \cdot c_d)}{\sum_{e' \in E} exp(E'^T_{e'} \cdot c_d)}$$



Outline

- ❑Introduction
  - ❑Motivation
  - ❑Related work
- ❑Our work
  - ❑Profile-matching
  - ❑Neural Network
- **❑Experiments**
- ❑Conclusion

Experiment Setup

- Socioeconomic Data and Applications Center (SEDAC)[1] Dataset
  - dataset citations in publications are manually labeled
- Experiments preparation
  - 849 publications are parsed on atmosphere research
  - 273 sections are identified to cite DOIs

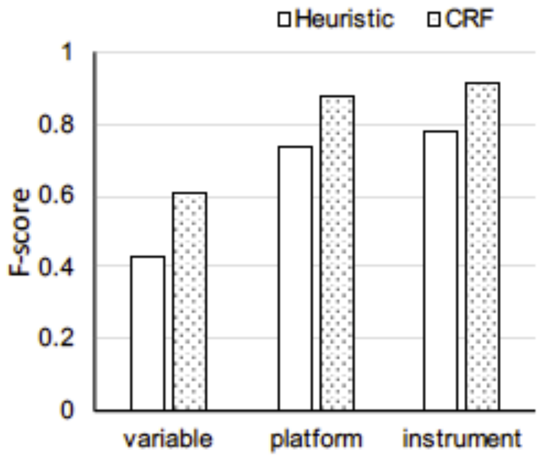
TABLE I: Source of Semantic Entities

Instrument	1,391
Platform	821
Variable	3,090
Data collection	41

1. Publicly available at http://sedac.ciesin.columbia.edu/citations-db

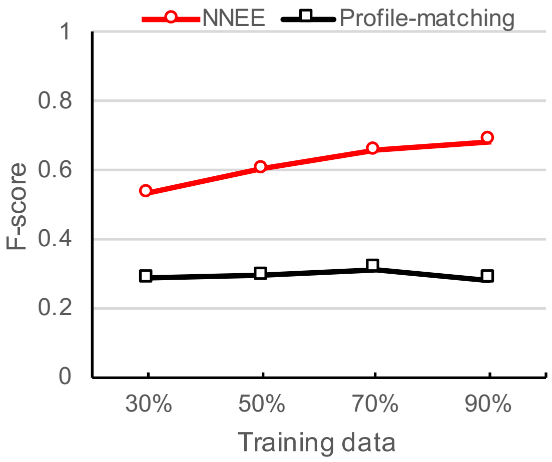
Explicit Entity Extraction Experiment

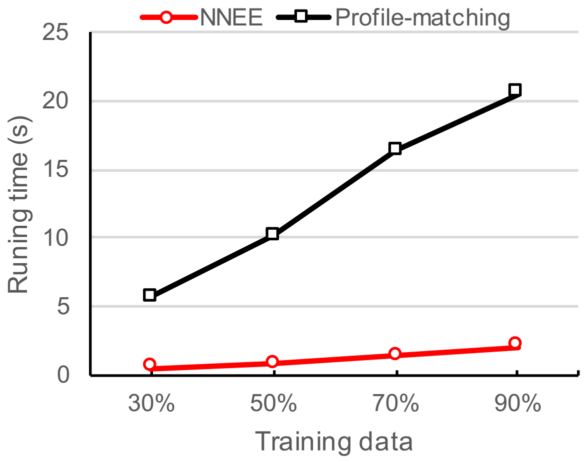
- Evaluation
  - Randomly select 14 papers to be evaluated by 5 domain experts
  - F-score: $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$
- Results
  - instruments and platforms are identified more accurate than variables
  - CRF model improves identification accuracy



Implicit Entity Extraction Experiment

- Ground truth: 273 sections with mentioned DOIs
- Results
  - Accuracy of NNEE is significantly higher than profile-matching method.
  - NNEE is fast and little increase as the amount of training data increases.





Outline

- ❑Introduction
  - ❑Motivation
  - ❑Related work
- ❑Our work
  - ❑Profile-matching
  - ❑Neural Network
- ❑Experiments
- ❑Conclusion

Conclusion

- Conclusion
  - simulate the cognitive process of how humans read articles
  - present NNEE to automatically extract semantic entities from unstructured academic papers
- Future work
  - model data analytics process
  - extend our approach to other research domains

Q&A

Thank you!