Information Extraction - 2

Zhiyuan Liu
liuzy@tsinghua.edu.cn
THUNLP

Outline

- Relation Extraction
  - Hand-written Patterns
  - Supervised RE
  - Semi-supervised RE
- Advanced Topics
  - Document-level RE
  - Open RE
  - Few-shot RE
  - Event Extraction

Relation Extraction

- Relation Extraction (RE)
  - Extract structured relational facts between two entities from text
  - Relational triple: (head entity, relation, tail entity)

Unstructured Text — Obama was born in Hawaii in 1961.

Named Entity Recognition ↓

Obama $_{\text{PER}}$ was born in Hawaii $_{\text{LOC}}$ in 1961 $_{\text{TIME}}$.

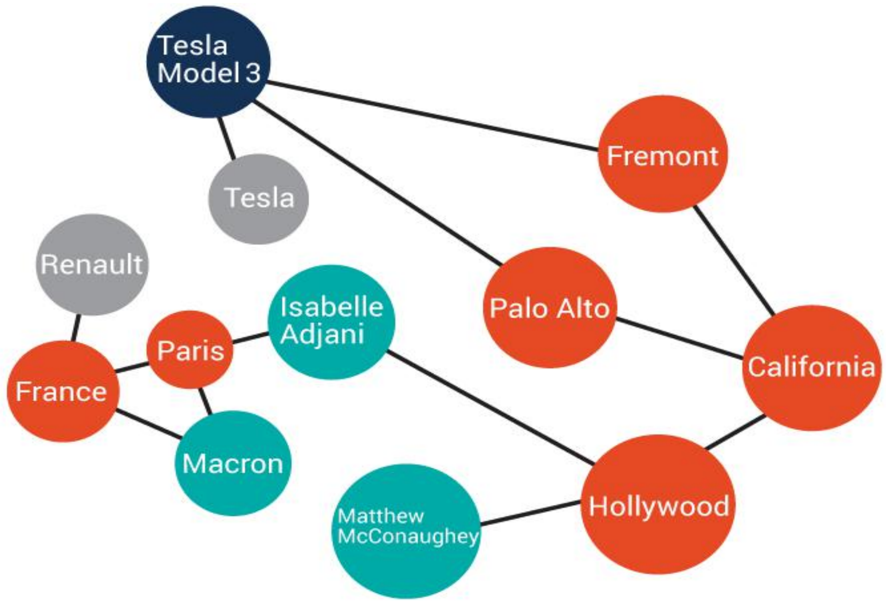
Relation Extraction ↓

Structured facts

(**Obama**, *Place of birth*, **Hawaii**)
(**Obama**, *Date of birth*, **1961**)

Why Relation Extraction?

- Knowledge Bases (KBs)
  - Contain a massive number of facts
  - Useful for many applications
  - However, KBs are still far from complete compared to the nearly infinite real-world facts



Why Relation Extraction?

- Supplement KBs via collaborative editing
    - Labor-intensive and time-consuming
    - World knowledge is ever-growing

- A huge number of relational facts are contained in unstructured and semi-structured texts from the Web
    - Encyclopedia
    - News articles
    - Academic articles
    - Blogs
    - …

Why Relation Extraction?

WIKIPEDIA
The Free Encyclopedia

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tsinghua University

From Wikipedia, the free encyclopedia

Coordinates: 40°00′00″N 116°19′36″E

Not to be confused with National Tsing Hua University.
"Tsinghua" redirects here. For other uses, see Tsinghua (disambiguation).

Tsinghua University (abbreviated **THU**; Chinese: 清华大学; also romanized as **Qinghua**) is a major research university in Beijing, and a member of the elite C9 League of Chinese universities.[10][11][12][13][14][15][*excessive citations*] Since its establishment in 1911, it has graduated numerous Chinese leaders in politics, business, academia, and culture.[16][17]

Reflecting its motto of *Self-Discipline and Social Commitment*, Tsinghua University is dedicated to academic excellence, advancing the well-being of Chinese society, and global development.[18] Tsinghua is

Tsinghua University

清华大学

Why Relation Extraction?

The New York Times

Why is Falcon Heavy different from other SpaceX rockets?

The company's workhorse is the Falcon 9 rocket, which first launched in 2010. The first stage of the Heavy essentially consists of three Falcon 9 first stages bound together. The second stages of the two rockets are identical.

The additional thrust allows the Heavy to propel 140,000 pounds to low-Earth orbit, nearly three times what the Falcon 9 can lift.

How does this Falcon Heavy differ from the first one?

On the test flight, the two side boosters were older versions reused from earlier flights. (SpaceX's best innovation to date is landing the booster stage of its rockets and launching it again; traditionally, rockets have been one-use throwaways, with the booster stages dropped into the ocean.)

For this one, the side boosters had never before been used. They were the latest version of the rocket, called "Block Five." ("Block" is what rocket companies call a major upgrade.) That boosts the thrust and how much the Falcon Heavy can carry.

Why Relation Extraction?

- Knowledge Bases (KBs) are far from complete
- Texts from the Web contain a huge number of relational facts

Motivates

- Relation Extraction (RE)
  - Automatically extract relational facts from text
  - Populate Knowledge bases

Relation Extraction

- How to build relation extraction systems?
  - Hand-written patterns
  - Supervised RE
  - Semi-supervised RE
    - Bootstrapping (using seeds)
    - Distant supervision

Hand-Written Patterns

- “Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use”
- What does ***Gelidium*** mean?
- How do you know?

Hand-Written Patterns

- "Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use”
- What does ***Gelidium*** mean?
- How do you know?

Hand-Written Patterns

- "Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use”
- Patterns for extracting IS-A relations
  - **Y, such as X**
  - such Y as X
  - X or other Y
  - X and other Y
  - Y including X
  - Y, especially X
  - ……

Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora. COLING 1992

Hand-Written Patterns

Pattern	Example
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	Bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara dang...
such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y, especially X	European countries, especially France, England, and Spain...

Hearst M A. Automatic Acquisition of Hyponyms from Large Text Corpora. COLING 1992

Relation Extraction

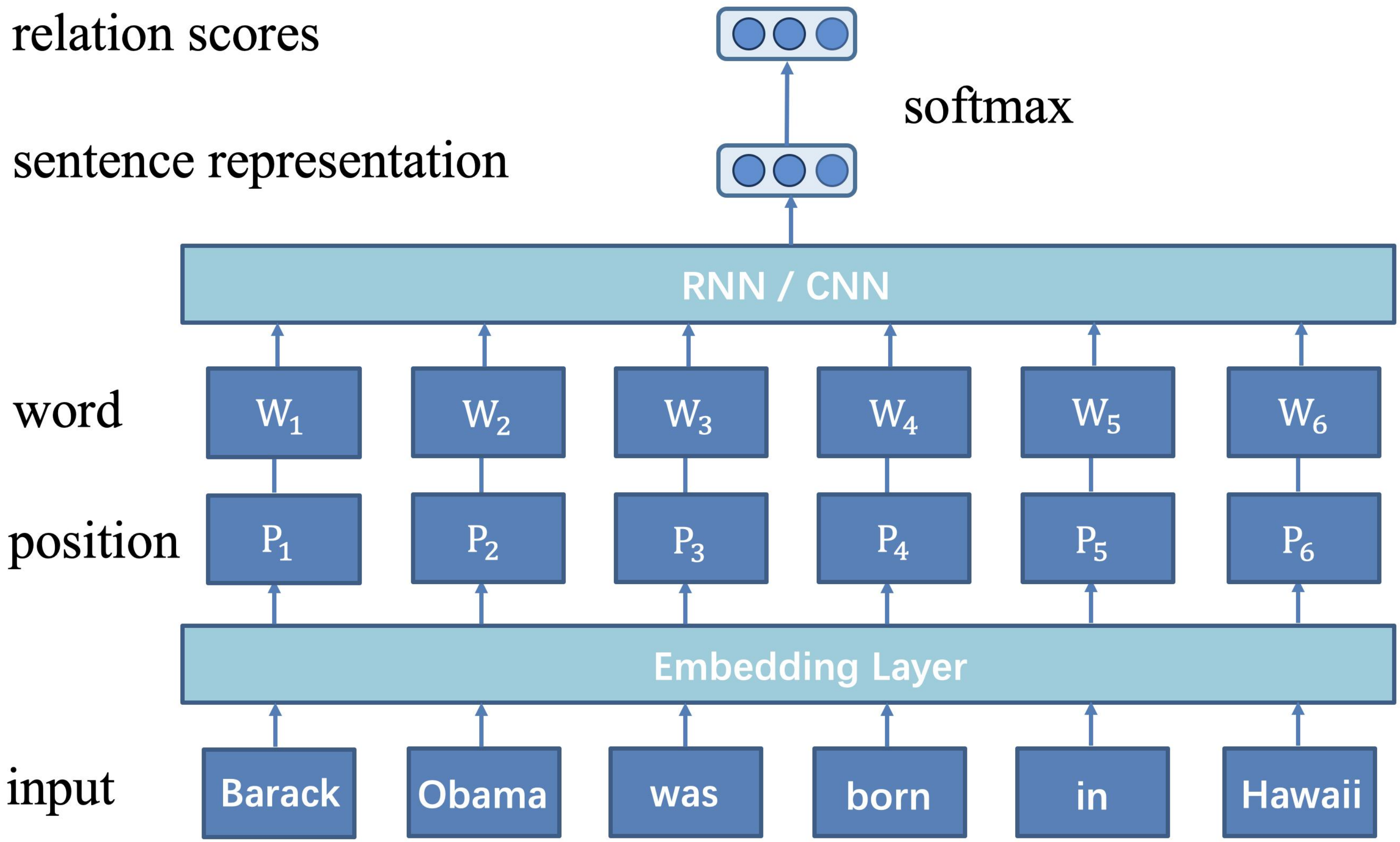
- How to build relation extraction systems?
  - Hand-written patterns
  - Supervised relation extraction
  - Semi-supervised
    - Bootstrapping (using seeds)
    - Distant supervision

Supervised Relation Extraction

- Choose a set of relations we would like to extract
- Choose a set of relevant named entities
- Select and label data
  - Choose a representative corpus
  - Label the named entities in the corpus
  - Hand-label the relations between these entities
- Train a relation classifier on the data

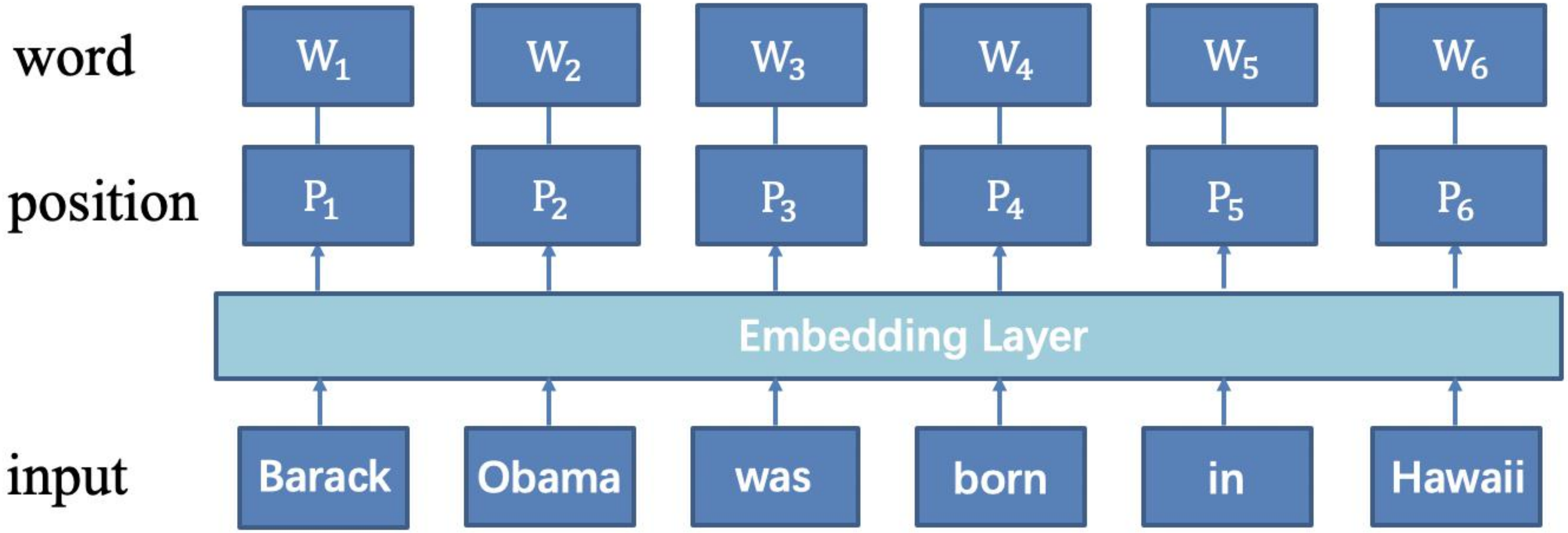
Supervised Relation Extraction

- RNN/CNN for supervised relation extraction



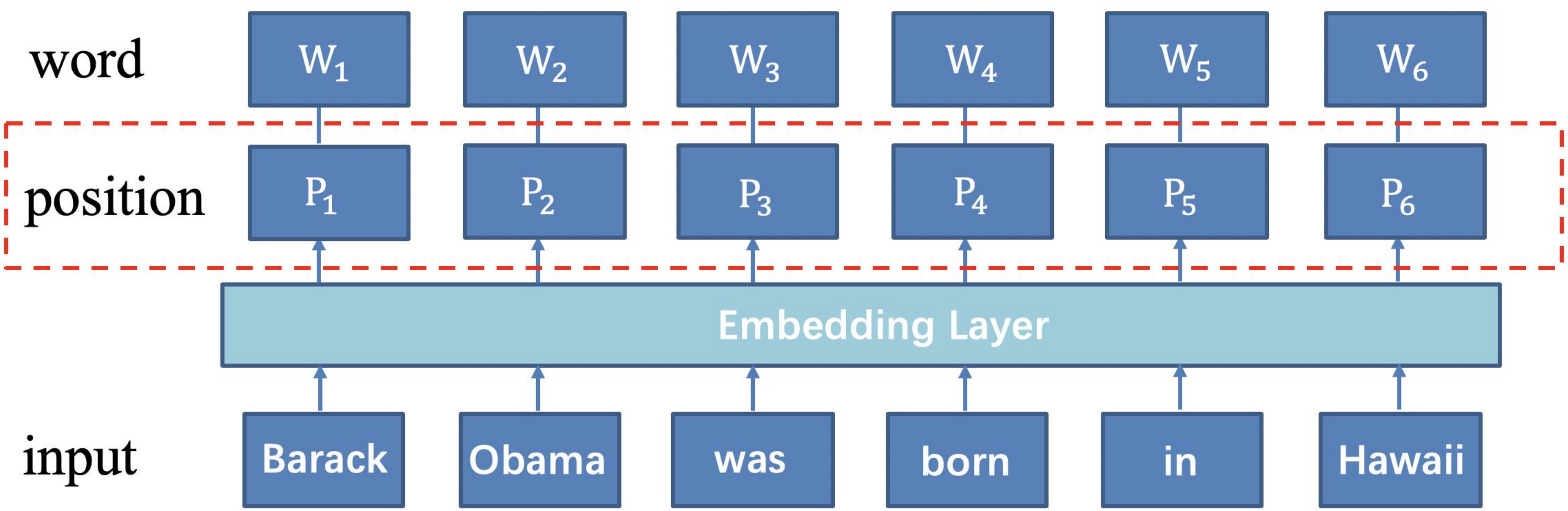
Supervised Relation Extraction

- Embedding Layer
  - Transform raw words of the sentence into low-dimensional vectors
  - Each word is represented as the concatenation of
    - Word embedding: encode word semantics
    - Position embedding: specify the position of entities



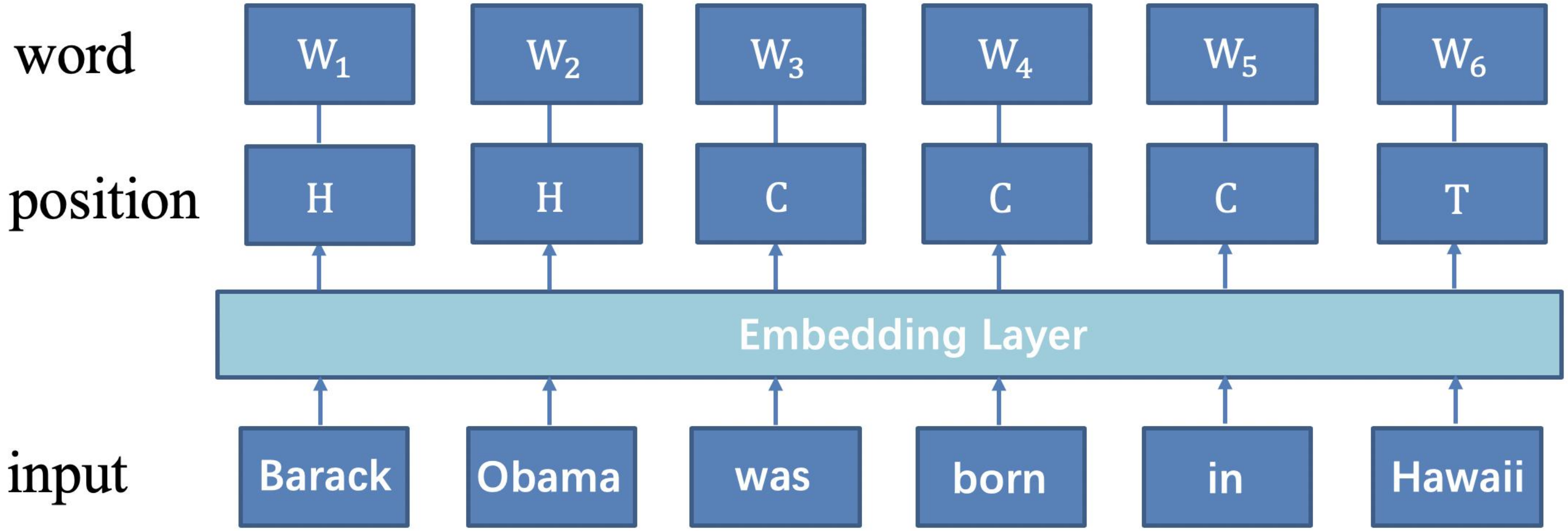
Supervised Relation Extraction

- Position embedding
  - Specify the position of entities
  - Several choices
    - Head-tail-context embedding
    - Relative distance embedding



Supervised Relation Extraction

- Head-tail-context embedding
  - Distinguish the words into
    - Head entity
    - Tail entity
    - Context words

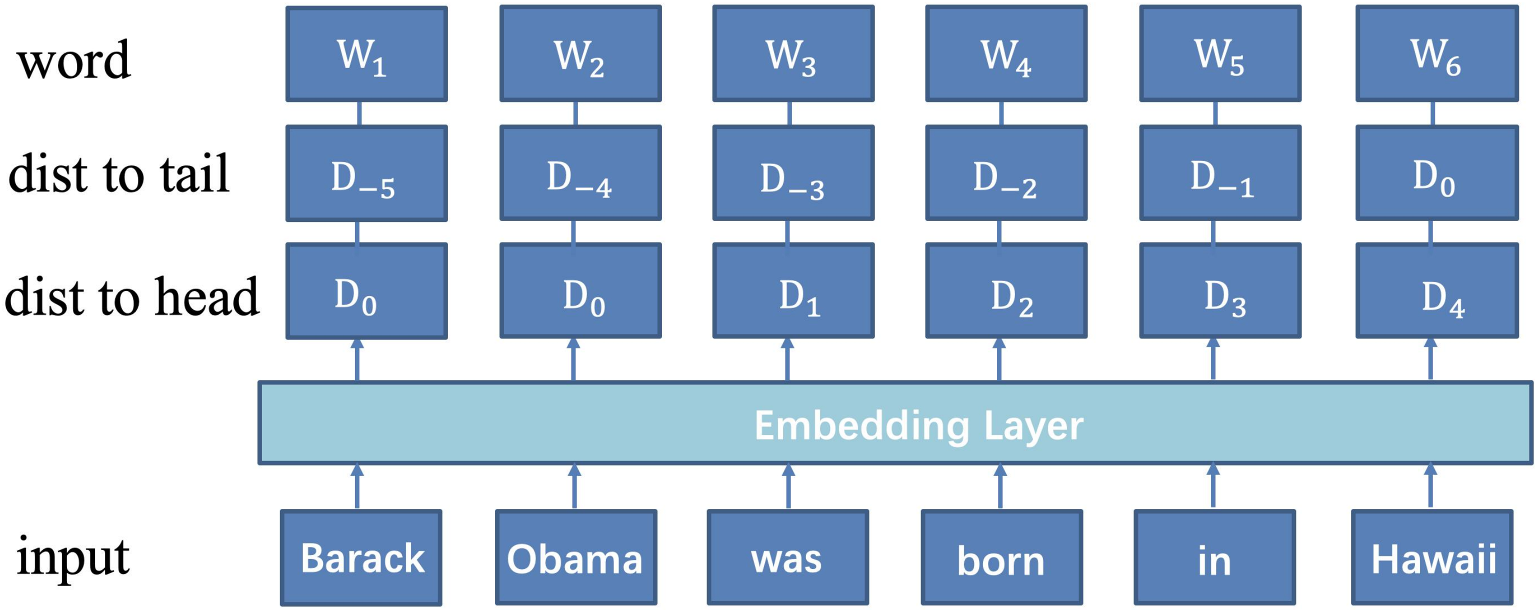


Supervised Relation Extraction

- Observation
  - Words close to the target entities are usually informative to determine the relation between entities
- Head-tail-context embedding
  - Agnostic to relative distance to entities
- Relative distance embedding
  - Encode the relative distance
  - Help RNN/CNN to keep track of how close each word is to head or tail entities

Supervised Relation Extraction

- Relative distance embedding is the concatenation of
  - The relative position to head entity
  - The relative position to tail entity

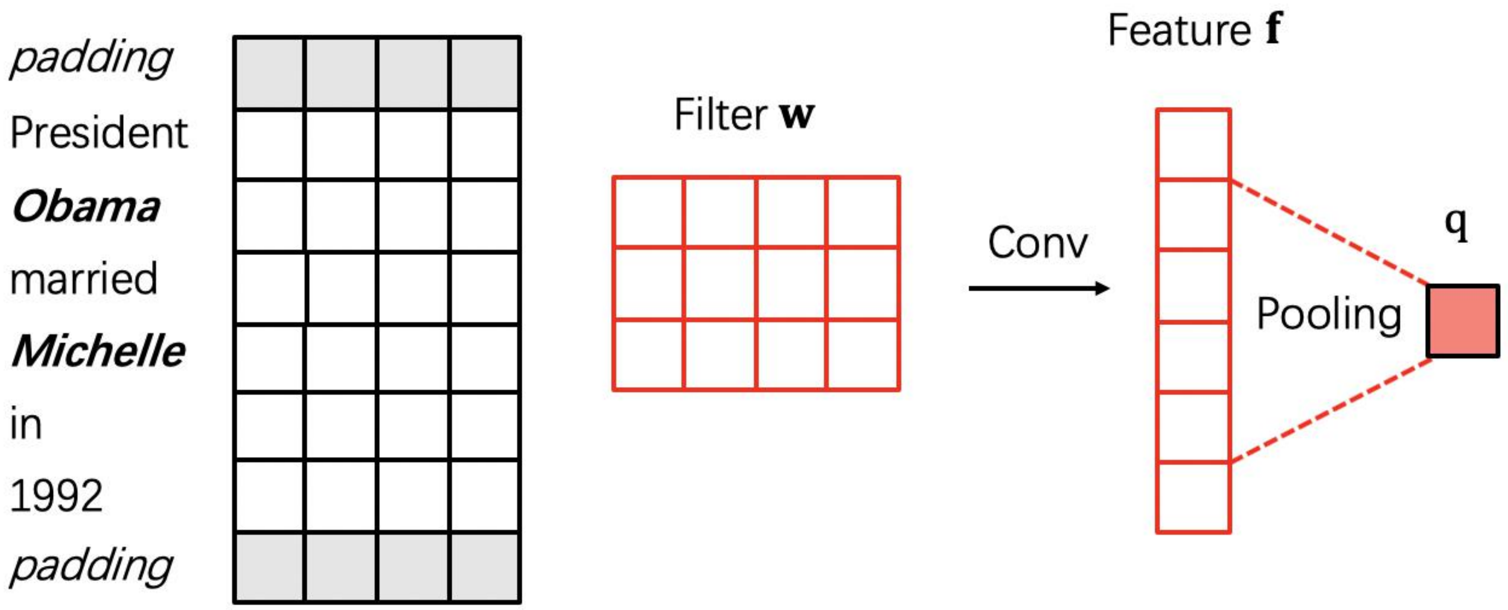


Supervised Relation Extraction

- RNN/CNN
  - Encode semantic information with respect to target entity pairs into sentence representation
- Softmax Layer
  - Transform sentence representation into probability distribution over predefined relation categories R
  - Note that R includes a special category-*NA*
    - *NA* indicates no relation between the target entity pair in the sentence
    - Or there is some kind of relation r between the target pair, but we are not interested, i.e. $r \notin R$

Convolutional Neural Networks

- Convolutional Neural Networks
  - Single max-pooling operation is utilized to determine the most significant features
  - Cannot capture the structural information between two entities



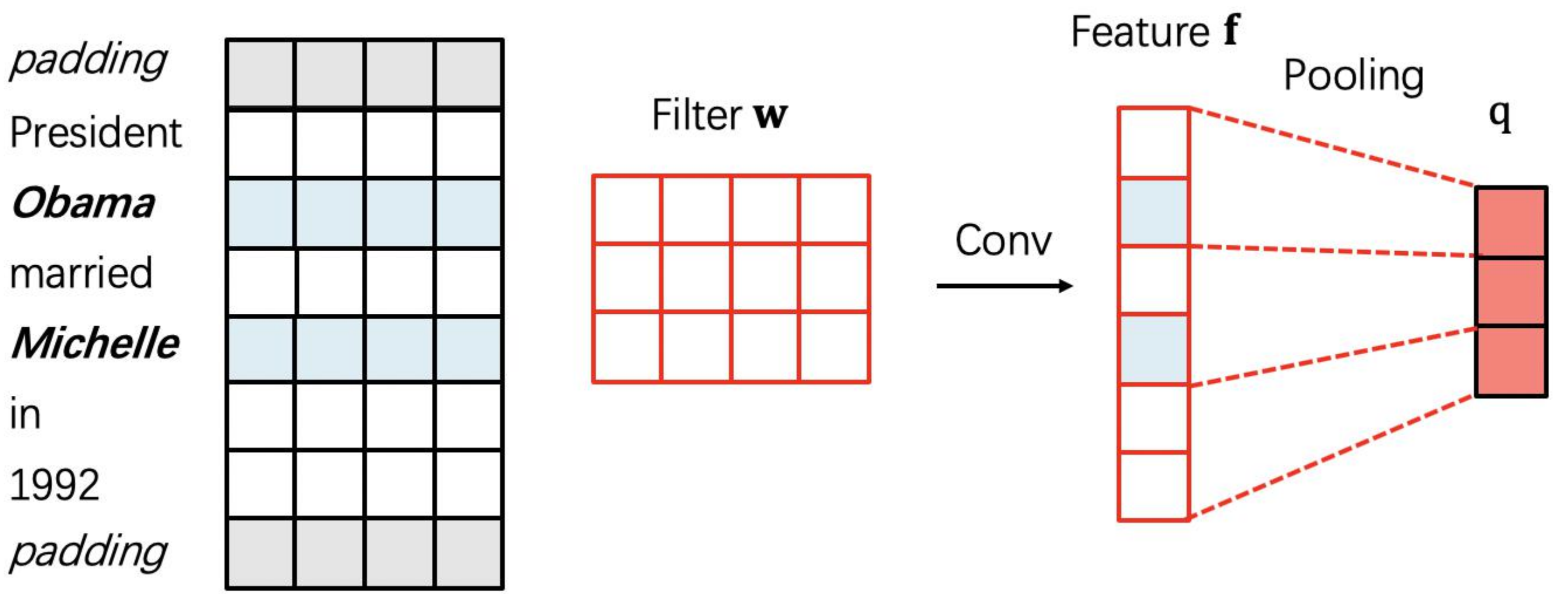
Piecewise Convolutional Neural Networks

- A sentence is inherently divided into three segments according to the two given entities
  - Internal context
    - characters inside the two entities
  - External context
    - characters around the two entities
  - Single max-pooling is not sufficient to capture such structural information

President ***Obama*** married ***Michelle*** in 1992

Piecewise Convolutional Neural Networks

- Piecewise Convolutional Neural Networks
  - Piecewise max-pooling layer
    - Divide the convolution results into three segments based on the positions of the two entities
    - Return the maximum value in each segment



Relation Extraction

- How to build relation extraction systems?
  - Hand-written patterns
  - Supervised relation extraction
  - Semi-supervised
    - Bootstrapping (using seeds)
    - Distant supervision

Semi-Supervised Relation Extraction

- Supervised approaches
  - Sentences in a corpus are hand-labeled for the presence relations between entities
  - Limitations
    - Labeling training data is labor-intensive
    - The scale of dataset is small, limiting the performance of relation extraction systems
    - Relations are labeled on a particular corpus. The resulting classifiers tend to be biased toward that text domain

Relation Bootstrapping

- No training set? Maybe you have
  - A few seed tuples
  - A few high-precision patterns
  - Large-scale unsupervised text data

- Can you use those seeds to do something useful?
  - Bootstrapping: Use the seeds to directly learn to populate a relation from unsupervised text data

Relation Bootstrapping

- Framework
  - Gather a set of seed entity pairs that have relation r
  - Iterate:
    1. Find sentences with these entity pairs
    2. Look at the context between or around the entity pair, and generalize the context to create patterns
    3. Use the patterns to obtain more relation instances and entity pairs for relation r from unsupervised text data

Brin S. Extracting patterns and relations from the world wide web. 1998

Relation Bootstrapping

- Example
  - Seed tuple: (Mark Twain, place of birth, Elmira)
    - "***Mark Twain*** was born in ***Elmira***, NY. "
    - "The hometown of ***Mark Twain*** is in ***Elmira***."
    - "***Elmira*** is ***Mark Twain***'s birthplace. "
  - Patterns for the relation:
    - X was born in Y
    - The hometown of X is in Y
    - Y is X's birthplace
  - Use those patterns to obtain new tuples
  - Iterate

Relation Bootstrapping

- Advantages
  - Use the immense amount of unlabeled text on the Web to create large-scale relation instances
  - Save heavy work-load of manually annotating relation instances

- Limitations
  - The resulting patterns often suffer from low precision and semantic drift

Relation Extraction

- How to build relation extraction systems?
  - Hand-written patterns
  - Supervised relation extraction
  - Semi-supervised
    - Bootstrapping (using seeds)
    - Distant supervision

Distant Supervision

- No training set? We have
  - Large-scale knowledge bases, such as Wikidata
  - Immense amount of unlabeled text on the Web
- Assumption of distant supervision
  - If two entities participate in a relation, any sentence that contains those two entities might express that relation
- Combine KBs, unlabeled text and distant supervision assumption
  - We can automatically create large-scale training data

Distant Supervision

- Framework
  - Use tuples in large-scale knowledge bases to get a huge number of seed examples
    - Potentially noisy, we will discuss it later
  - Train a supervised classifier on the resulting data

- Advantages
  - An effective method of automatically labeling training data
  - Can combine information from many different instances of the same relation
  - Do not require iteratively expanding patterns

Distant Supervision

- Rethink the assumption of distant supervision
  - If two entities participate in a relation, any sentence that contains those two entities might express that relation
  - This assumption is too strong and could result in wrong labeling problem

- Consider the following example
  - For a tuple (Obama, spouse, Michelle)
  - “***Obama*** is three years older than ***Michelle***”
  - The instance does not express the relation *spouse*

Distant Supervision

- How much noise does distant supervision introduce?
  - Manual verification shows that approximately 80% of relation instances are correctly labeled with distant supervision
  - The wrong labeling problem hinders the performance of a model trained on such noisy data

Sorokin D, Gurevych I. Context-aware representations for knowledge base relation extraction. EMNLP 2017.

Distant Supervision

- To alleviate the wrong-labeling problem
  - Distantly supervised relation extraction can be treated as a multi-instance problem
- Multi-instance problem
  - Data organization
    - The training set consists of many bags
    - Each bag contains many instances
  - Label
    - The labels of the bags are known
    - The labels of the instances in the bags are unknown
  - Goal
    - Predict the labels of the unseen bags

Distant Supervision

- Multi-instance problem

Training

bag label: place of birth

(Mark Twain, Elmira)

"*Mark Twain* was born in *Elmira*, NY." **instance label: ?**

"*Mark Twain* went to *Elmira* himself." **instance label: ?**

"*Elmira* is *Mark Twain*'s birthplace." **instance label: ?**

Testing

bag label: ?

(Obama, Michelle)

"*Obama* married *Michelle* in 1992." **instance label: ?**

"*Obama* is the husband of *Michelle*." **instance label: ?**

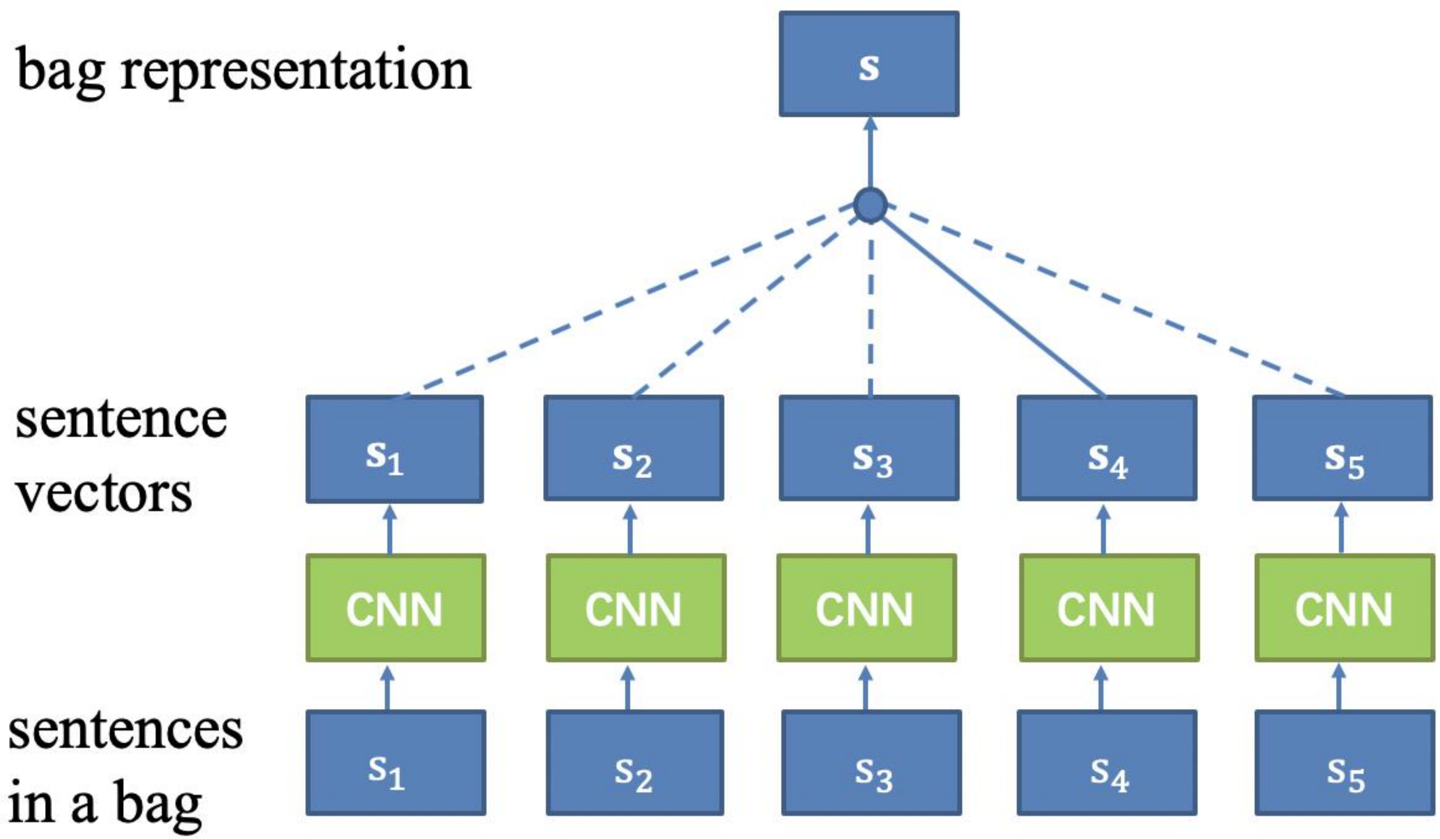
"*Obama* is older than *Michelle*." **instance label: ?**

Distant Supervision

- Multi-instance problem
  - Instead of supervising the relation classifier at the instance level
    - Since we do not know the exact label of each instance
  - Design objective function at the bag level
    - The labels of bags are known
  - Take the uncertainty of instance labels into account

Distant Supervision

- At-least-one multi-instance learning
  - Select the most likely sentence for each entity pair



Distant Supervision

- At-least-one multi-instance learning
  - A bag is labeled as r if and only if
    - At least one instance in the bag is assigned as r
  - Design objective function at the bag level

$$J(\theta) = \sum_{i=1}^{T} \log P(y_i | s_i^j; \theta)$$

- T is the number of bags, θ is trainable parameters
- y_i is the label of the i-th bag
- s_i^j is the instance with highest score for y_i in the bag S_i

$$j = \arg\max_{q \in B_i} P(y_i | s_i^q; \theta)$$

Distant Supervision

- At-least-one multi-instance learning
  - Assume that at least one sentence that mentions two entities will express their relation
  - Only select the most likely sentence for each entity pair in training and prediction
  - Lose rich information in neglected sentences

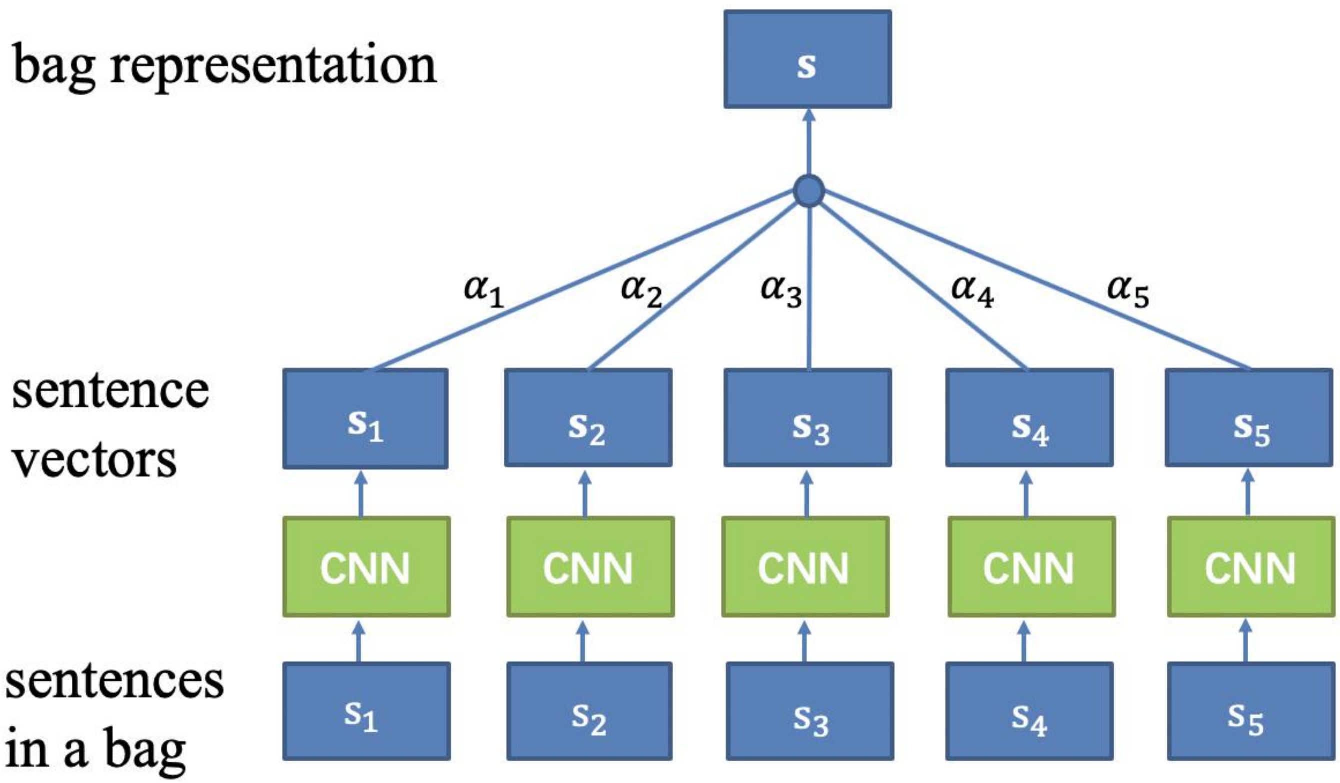
Distant Supervision

- Sentence-level attention over multi-instances
  - Goal
    - Utilize all informative sentences in the bag
  - Method
    - Instead of only selecting the most likely sentence
    - Use sentence-level attention over multiple instances
    - Represent the relation as semantic composition of sentence embeddings
    - Dynamically reduce the weights of those noisy instances

Lin Y, et al. Neural relation extraction with selective attention over instances. ACL 2016.

Distant Supervision

- Sentence-level attention over multi-instances
  - Represent the relation as weighted sum of sentence embeddings via attention



Distant Supervision

- Sentence-level attention over multi-instances
  - Represent the relation as weighted sum of sentence embeddings $\{\mathrm{s}_1, \mathrm{s}_2, \ldots, \mathrm{s}_n\}$

$$\mathbf{s} = \sum_i \alpha_i \mathbf{s}_i$$

    - α_i is the weight of instance s_i

$$\alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)}$$

    - e_i scores how well the input sentence s_i and the predicted relation r matches

Distant Supervision

- Sentence-level attention over multi-instances
  - e_i scores how well the input sentence s_i and the predicted relation r matches
  - e_i is obtained via a bilinear function

$$e_i = \boldsymbol{s}_i \mathbf{A} \mathbf{r}$$

- $\mathbf{A}$ is a learnable weighted diagonal matrix
- $\mathbf{r}$ is the query vector of relation r

Distant Supervision

- Sentence-level attention over multi-instances
  - The scores $\mathbf{o}$ for all relations are obtained via a linear transformation

$$\mathbf{o} = \mathbf{M}\mathbf{s} + \mathbf{d}$$

  - The probability of a relation r for a bag S is given by a softmax layer

$$\mathrm{P}(r|S) = \frac{\exp(o_r)}{\sum_k \exp(o_k)}$$

Summary of Distant Supervision

- Distant supervision
  - Align large-scale knowledge bases and unlabeled text
  - An effective method of automatically labeling training data
  - Could result in wrong labeling problem
- Alleviate wrong labeling problem
  - At-least-one multi-instance learning
    - Select the most likely sentence for each entity pair
    - Take the uncertainty of instance labels into account
    - Lose rich information containing in neglected sentences
  - Sentence-level attention over multi-instances
    - Represent the relation as semantic composition of sentence embeddings via attention
    - Dynamically reduce the weights of those noisy instances

Outline

- Relation Extraction
  - Hand-written Patterns
  - Supervised RE
  - Semi-supervised RE
- Advanced Topics
  - Open RE
  - Document-level RE
  - Few-shot RE
  - Event Extraction

Open Relation Extraction

- Goal
  - Aim to discover new relation types from unsupervised open-domain corpora.
- Definition
  - Given a sentence and two entities mentioned in the the sentence
  - Extract the phrases which represent the relation

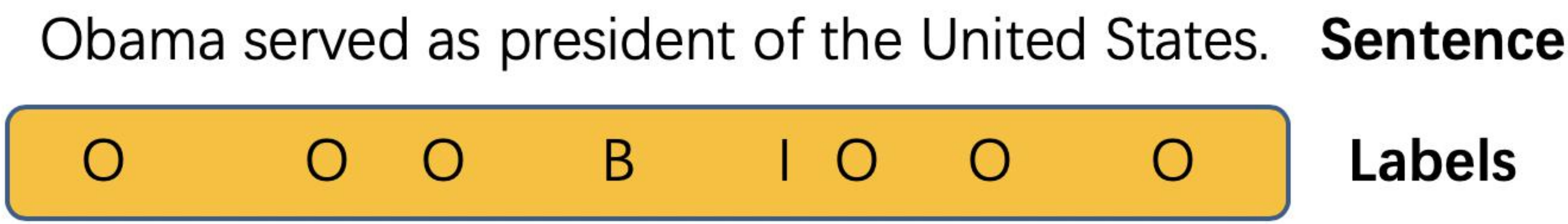
Text: From 2009 to 2017, Obama served as president of the United States.
Entity 1: Obama
Entity 2: the United States
Extraction Result: president of

Open Relation Extraction

- Methods
  - Tagging-based model
    - Cast OpenRE as a sequence labeling problem.
  - Clustering-based model
    - Cluster semantic patterns into several relation types.

Open Relation Extraction

- Tagging-based methods
  - Label each token in the sentence with BIO or BIOES labels.



- BIO
  - B: Beginning of the phrase
  - I: Inside the phrase
  - O: Outside the phrase

Open Relation Extraction

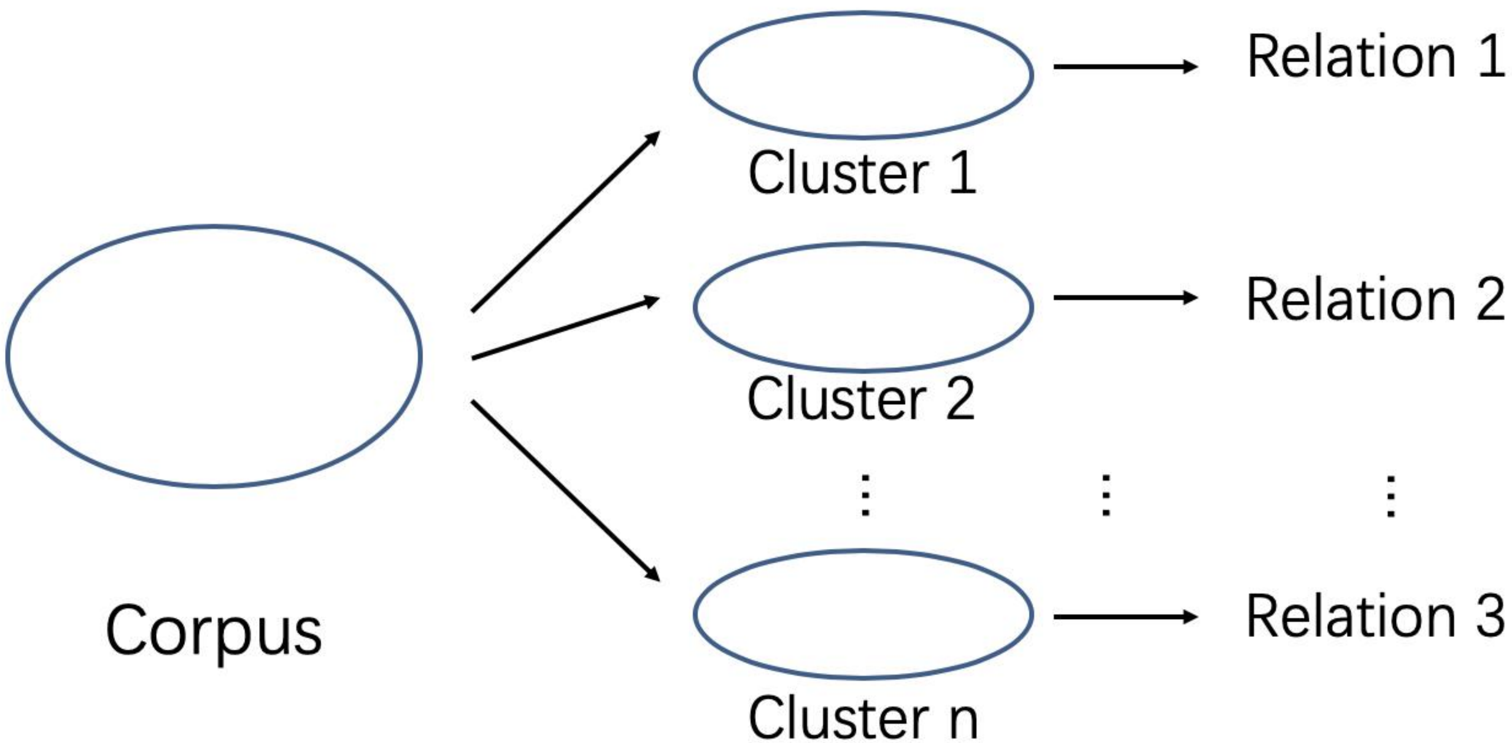
- Limitations of tagging-based methods
  - Some relations cannot be explicitly represented with tokens in sentences
    - ***Obama*** and ***Michelle*** have two daughters. (spouse)

  - It is hard to align different relational tokens that have the same meanings
    - ***Obama*** graduated from ***Harvard University***.

      Obama was educated at ***Harvard University***.

Open Relation Extraction

- Clustering-based methods
  - The sentences in the same cluster share the same relation.



Outline

- Relation Extraction
  - Hand-written Patterns
  - Supervised RE
  - Semi-supervised RE
- Advanced Topics
  - Open RE
  - Document-level RE
  - Few-shot RE
  - Event Extraction

Document-Level Relation Extraction

- Sentence-level relation extraction
  - Head and tail entities must come from one sentence
  - Relations are limited in single sentences

- According to the data samples from Wikipedia, a large number (42.2%) of relational facts are expressed in multiple sentences

Yao, et al. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. ACL 2019

Document-Level Relation Extraction

- Document-level relation extraction
  - Extract relational facts from multiple sentences in a document

Kungliga Hovkapellet

[1] ***Kungliga Hovkapellet*** (The ***Royal Court Orchestra***) is a *Swedish* orchestra, originally part of the *Royal Court* in ***Sweden***'s capital *Stockholm*. [2] The orchestra originally consisted of both musicians and singers. [3] It had only male members until *1727*, when *Sophia Schröder* and *Judith Fischer* were employed as vocalists; in the *1850s*, the harpist *Marie Pauline Åhman* became the first female instrumentalist. [4] From *1731*, public concerts were performed at ***Riddarhuset*** in *Stockholm*. [5] Since *1773*, when the ***Royal Swedish Opera*** was founded by *Gustav III* of ***Sweden***, the ***Kungliga Hovkapellet*** has been part of the opera's company.

Subject: *Kungliga Hovkapellet; Royal Court Orchestra*
Object: *Royal Swedish Opera*
Relation: `part_of` **Supporting Evidence:** 5

Subject: *Riddarhuset*
Object: *Sweden*
Relation: `country` **Supporting Evidence:** 1, 4

Document-Level Relation Extraction

- Document-level relation extraction requires different kinds of complex reasoning skills

Reasoning Types	%	Examples
Pattern recognition	38.9	[1] ***Me Musical Nephews*** is a **1942** one-reel animated cartoon directed by Seymour Kneitel and animated by Tom Johnson and George Germanetti. [2] Jack Mercer and Jack Ward wrote the script. ... **Relation:** `publication_date` **Supporting Evidence: 1**
Logical reasoning	26.6	[1] "Nisei" is the ninth episode of the third season of the American science fiction television series The X-Files. ... [3] It was directed by David Nutter, and written by **Chris Carter**, Frank Spotnitz and Howard Gordon. ... [8] The show centers on FBI special agents ***Fox Mulder*** (David Duchovny) and Dana Scully (Gillian Anderson) who work on cases linked to the paranormal, called X-Files. ... **Relation:** `creator` **Supporting Evidence: 1, 3, 8**
Coreference reasoning	17.6	[1] ***Dwight Tillery*** is an American politician of the Democratic Party who is active in local politics of Cincinnati, Ohio. ... [3] He also holds a law degree from the **University of Michigan Law School**. ***Tillery*** served as mayor of Cincinnati from 1991 to 1993. ... **Relation:** `educated_at` **Supporting Evidence: 1, 3**
Common-sense reasoning	16.6	[1] ***William Busac*** (1020-1076), son of William I, Count of Eu, and his wife Lesceline. ... [4] ***William*** appealed to King Henry I of France, who gave him in marriage **Adelaide**, the heiress of the county of Soissons. [5] **Adelaide** was daughter of Renaud I, Count of Soissons, and Grand Master of the Hotel de France. ... [7] ***William*** and **Adelaide** had four children: ... **Relation:** `spouse` **Supporting Evidence: 4, 7**

Outline

- Relation Extraction
  - Hand-written Patterns
  - Supervised RE
  - Semi-supervised RE
- Advanced Topics
  - Open RE
  - Document-level RE
  - Few-shot RE
  - Event Extraction

Few-Shot Learning

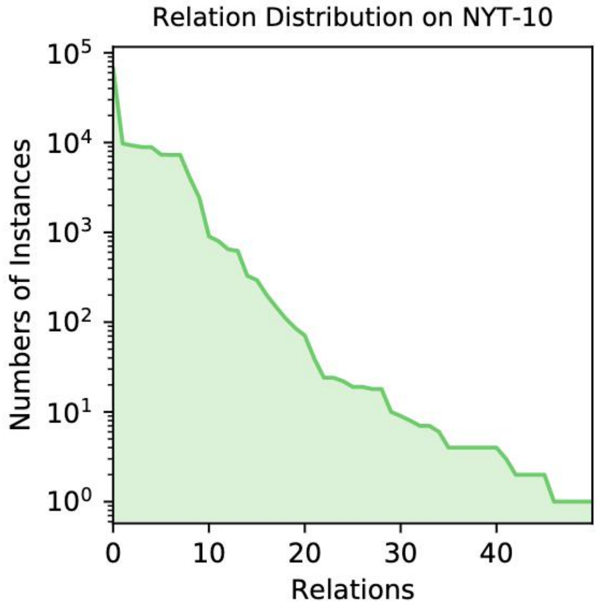
- Learn new tasks with only a handful of examples
  - Deep learning models usually require large amounts of data to achieve promising results
  - Human can grasp new knowledge with only a few samples.
  - Can machine learn like human beings?

Few-Shot Learning

- Why do we need few-shot learning
  - A playground for AI models
  - Reduce the heavy reliance on large-scale data
  - For domains where data is hard to collect

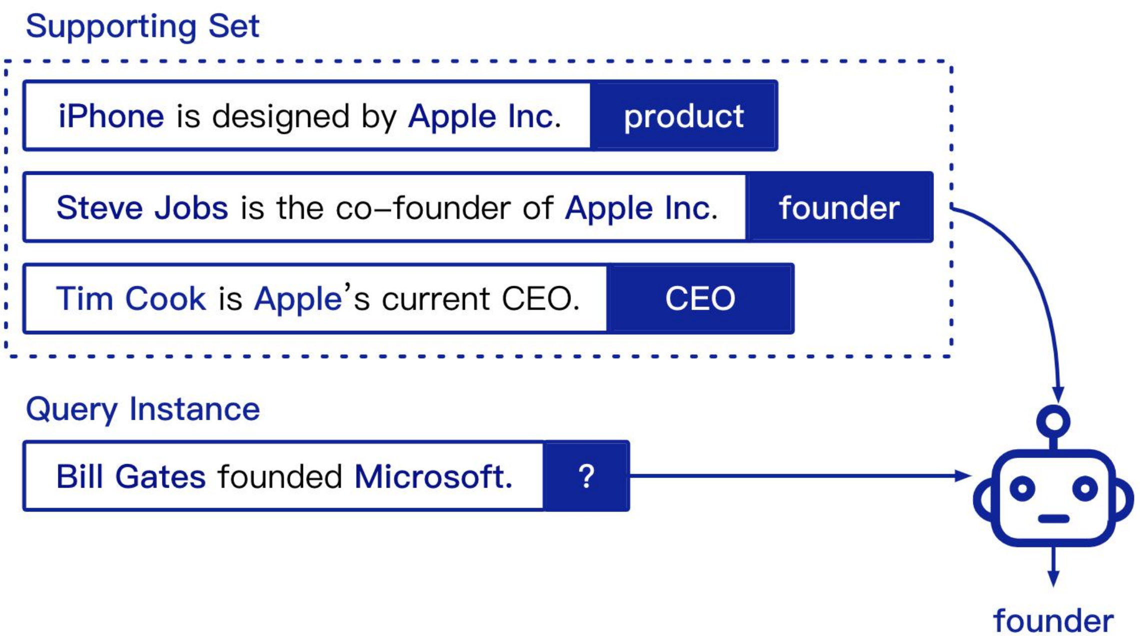
Few-Shot for Relation Extraction

- Why do we need few-shot learning for RE
  - Relational data suffer from the long-tail problem
    - Many relations have less than 10 instances in distant supervision dataset NYT-10.



Few-Shot for Relation Extraction

- Why we need few-shot learning for RE
  - We need learning methods with higher efficiency, so that they can learn new relations with only a few examples



Few-Shot for Relation Extraction

- **Datasets**
  - FewRel [1] and FewRel 2.0 [2]
    - A large-scale few-shot relation extraction dataset
- **Methods**
  - Key: transfer learning
  - Metric learning methods
    - E.g., Prototypical Networks [3]
  - Meta learning methods
    - E.g., Model-Agnostic Meta-Learning [4]
- We will give details in later class

[1] FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. Han et al. EMNLP 2018.
[2] FewRel 2.0: Towards More Challenging Few-Shot Relation Classification. Gao et al. EMNLP 2019.
[3] Prototypical Networks for Few-shot Learning. Snell et al. NIPS 2017.
[4] Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Finn et al. ICML 2017.

Outline

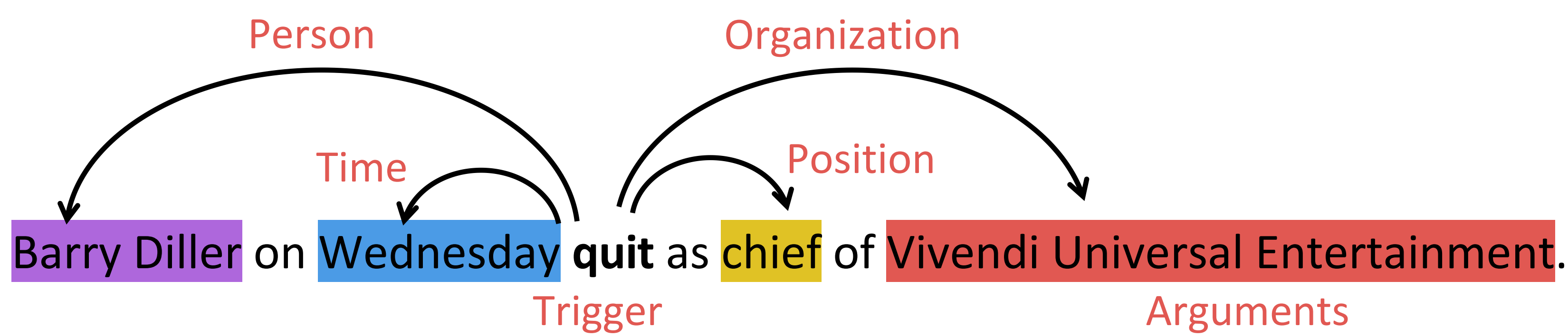
- Relation Extraction
  - Hand-written Patterns
  - Supervised RE
  - Semi-supervised RE
- Advanced Topics
  - Open RE
  - Document-level RE
  - Few-shot RE
  - Event Extraction

Event Extraction: Outline

- Introduction
  - Task definition
  - Challenges
- Supervised methods
- Distant supervision methods

Event Extraction: Definition

- To extract structural event knowledge from plain text.
- Predict event triggers with specific event types and arguments for each trigger.



Event Extraction: Terminology

- **Event mention:** a phrase or sentence within which an event is described, including a trigger and arguments.
- **Event trigger:** the main word that most clearly expresses the occurrence of an event (typically a verb or a noun).
- **Event argument:** an entity mention, temporal expression or value (e.g., job-title) that is involved in an event (i.e., participants).
- **Argument role:** the relationship between an argument to the event in which it participates.

Event Extraction: Pipeline

- Event Detection (ED): detect event triggers and classify the corresponding event types

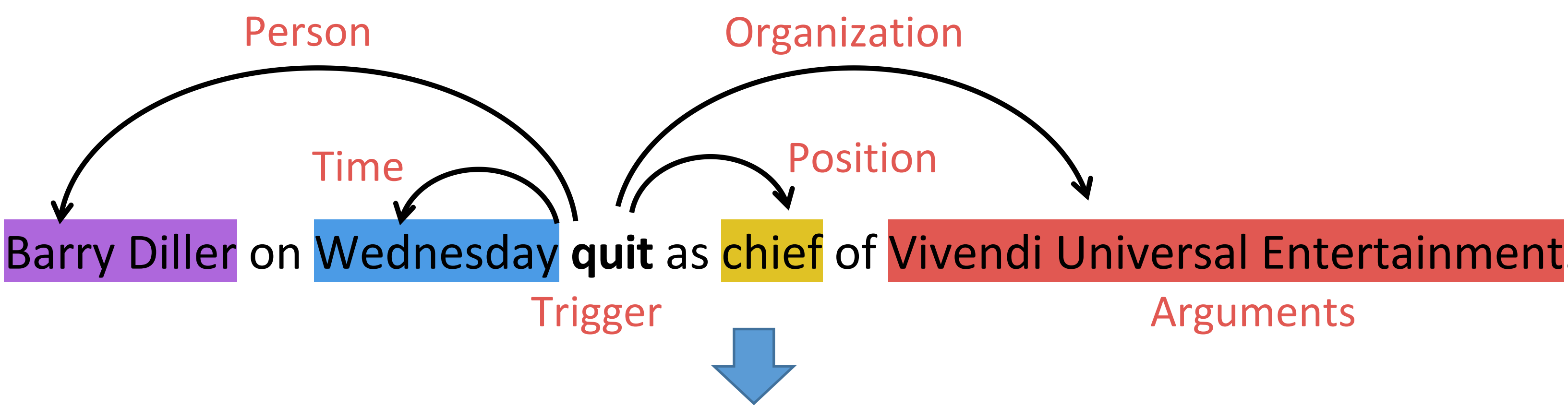
Barry Diller on Wednesday **quit** as chief of Vivendi Universal Entertainment.

Trigger

Trigger	**Quit(a "Personnel/End-Position" event)**

Event Extraction: Pipeline

- Event Argument Extraction (EAE): find the entities which serve as event arguments for the detected event, and classify the argument roles



Trigger	**Quit(a “Personnel/End-Position” event)**	
Arguments	Role = Person	Barry Diller
	Role = Organization	Vivendi Universal Entertainment
	Role = Position	chief
	Role = Time	Wednesday

Event Extraction: Challenges

- Complex structure:
  - Event structures are much more complex than entity relations
  - Inherent dependencies: trigger-argument, argument-argument
- Data sparsity:
  - ACE 2005 English: 599 documents, 6000 labeled sentences
  - Hard to annotate

Supervised Methods

- Feature-based model
- Neural model
- To model complex structures

Feature-based Model

- Human-designed features
  - Lexical-level features
    - Part-of-speech tags
    - Entity information

Event Type

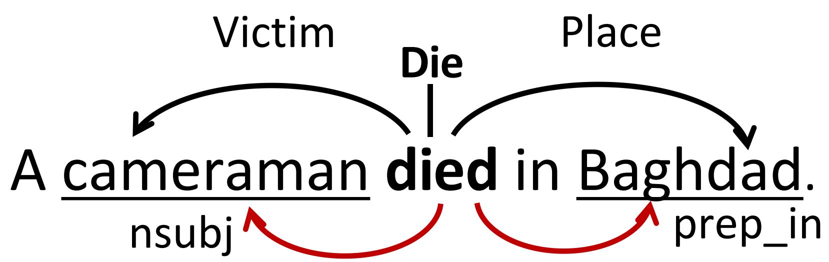
He left the **bathroom**. → Transport

He left **Microsoft**. → End-Position

Joint Event Extraction via Structured Prediction with Global Features. Li et al. ACL 2013.

Feature-based Model

- Human-designed features
  - Sentence-level features
    - Dependency parsing



nsubj -> (cameraman plays the Victim role in die event)

nsubj: A nominal subject which is the syntactic subject of a clause.

Feature-based Model

- Human-designed features
  - Document-level features
    - Cross-event Inference

Event Type

They held a party for his retirement.

He planned to go shopping.

He **left** the company.

End-Position

Transport

Feature-based Model

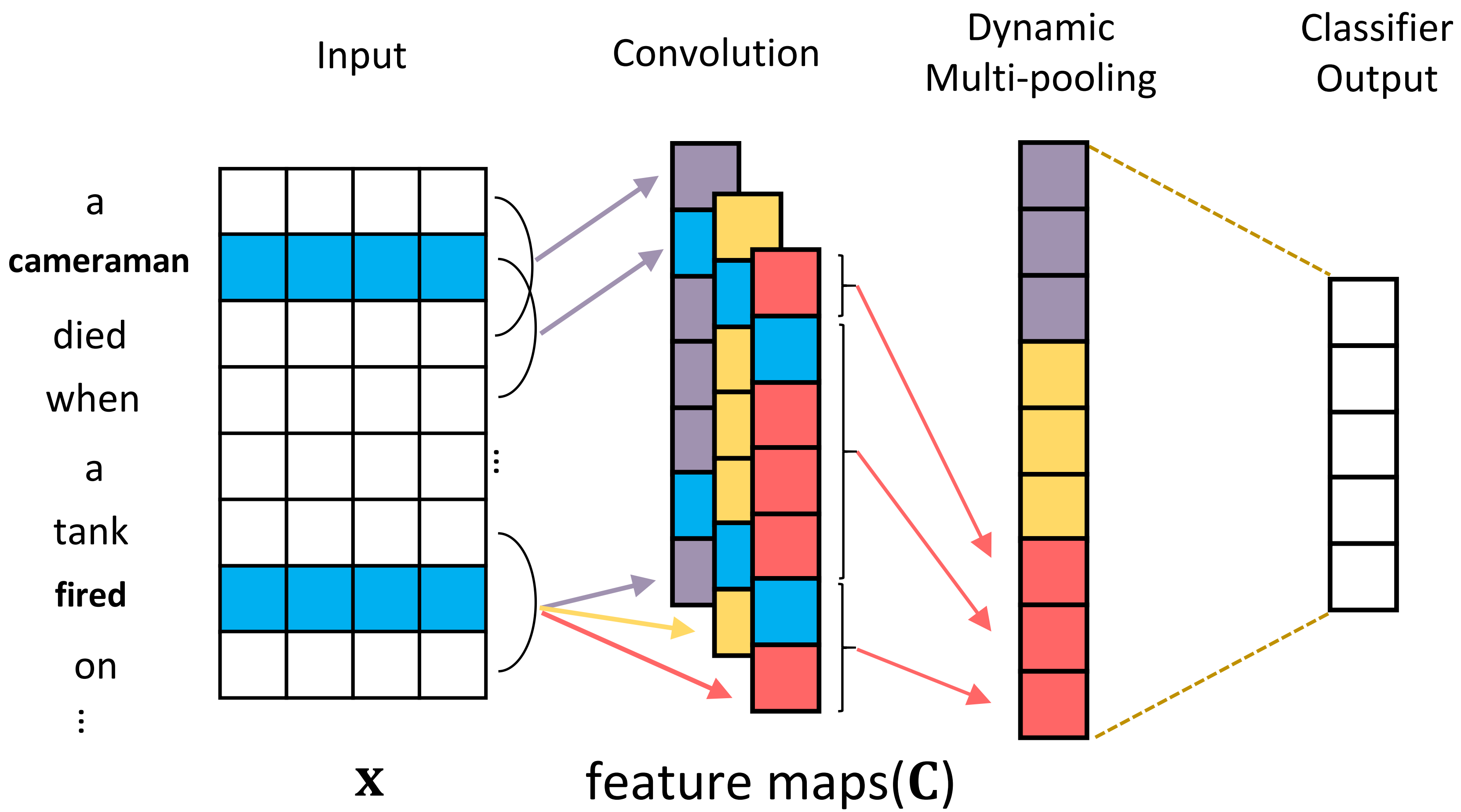
- Human-designed features
  - Lexical-level features
  - Sentence-level features
  - Document-level features
- Rely heavily on NLP tools for feature extraction
  - Introduce parsing errors and lead to error propagation, especially for low-resources languages

Neural Model: DMCNN

- One of the earliest model to adopt neural network in EE
- Automatically learn effective lexical-level and sentence-level features from plain text.
- DMCNN Architecture
  - Input Layer
  - Convolutional Neural Network
  - Dynamic Multi-pooling Layer
  - Classifier Output Layer

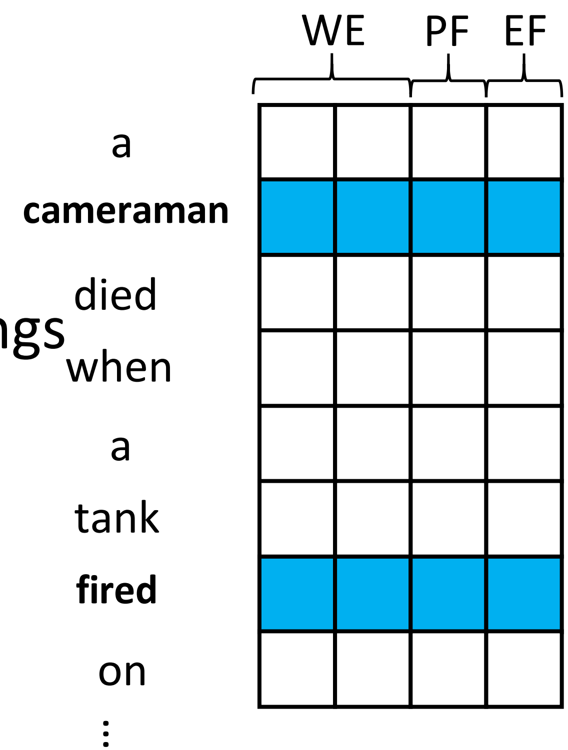
Event extraction via dynamic multi-pooling convolutional neural networks. Chen et al. ACL 2015.

DMCNN Architecture



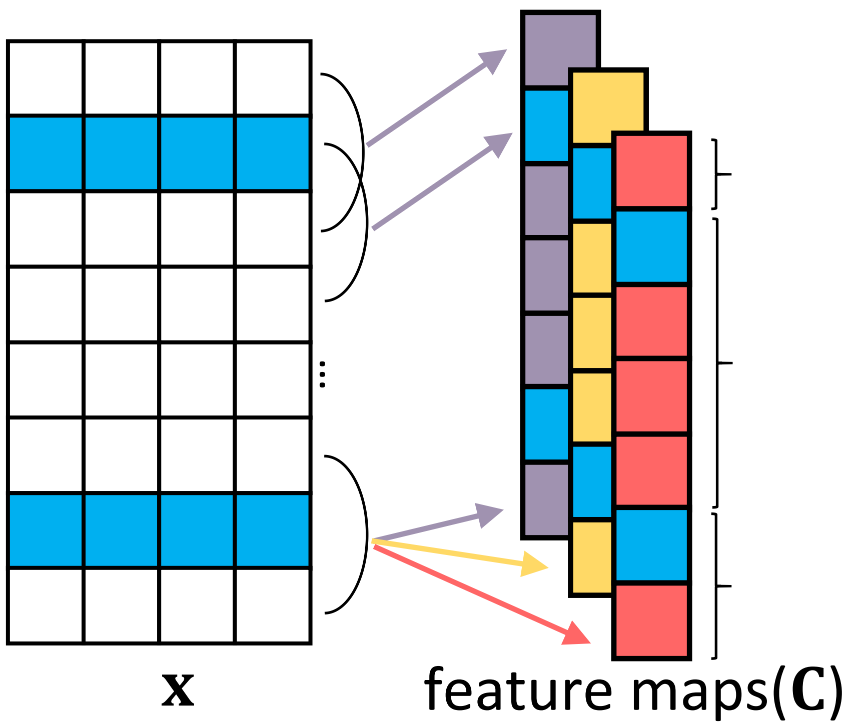
DMCNN: Input Layer

- Transform words into input representations $\mathbf{x}$ via word embeddings and event-specific features
- $\mathbf{x} \in \mathbb{R}^{n \times (d + d_p + d_e)}$: input representation
  - n is the length of sentence
  - d is the dimension of word embeddings
  - d_p is the dimension of position embeddings
  - d_e is the dimension of event type embeddings
- WE: word embeddings
- PF: Position features
- EF: Event-type features



DMCNN: Convolution Layer

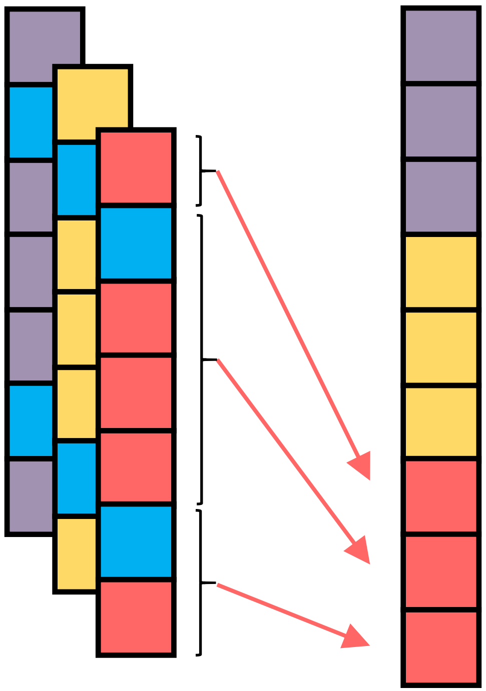
- Capture the semantics of the entire sentence into the feature map $\mathbf{C} \in \mathbb{R}^{m*(n-h+1)}$ using multiple filters.
  - m is the number of filters
  - h is the size of a window
  - $c_{ji} = \tanh\left(\omega_j \cdot x_{i:i+h-1} + b_j\right)$



DMCNN: Dynamic Multi-pooling

- Extract event-oriented features within feature maps.
  - Split each feature map into three parts according to positions of the argument candidate and the trigger
  - Perform pooling within each part then concatenate. (Similar to PCNN)

A **cameraman** died when a tank **fired** on …



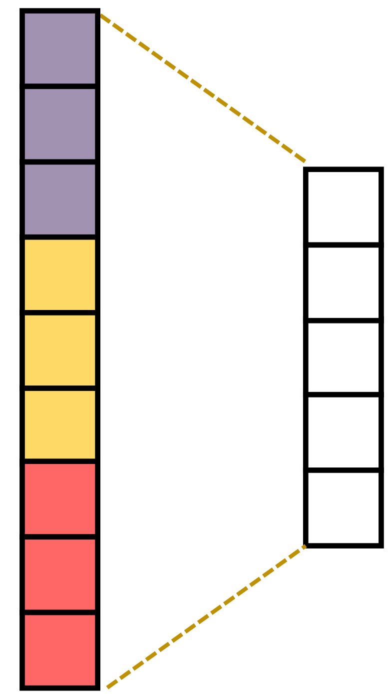
feature maps

DMCNN: Classifier Output

- View event extraction as a multi-class classification task
- Feed the feature vector into a classifier to compute the confidence of each argument role

$$\boldsymbol{O} = \mathrm{softmax}(\boldsymbol{W} \cdot \boldsymbol{F} + b_s)$$

  - $\boldsymbol{W}$ and b_s are the trainable parameters



DMCNN: Summary

- Leveraging Neural Network
  - Avoid complicated feature engineering
  - Better generalization ability
  - Better performance (2%-3% improvement in F1)
- Dynamic Multi-Pooling
  - An easy way to aggregate the extraction-oriented information in the sentence
  - Cannot well model the complex structures
- How to model the structures (dependencies) ?
  - HMEAE: Hierarchical Modular Event Argument Extraction. Wang et al. EMNLP 2019.
  - Graph convolutional networks with argument-aware pooling for event detection. Nguyen et al. AAAI 2018.

Distant Supervision Methods

- It’s hard to construct large-scale EE datasets with human efforts
  - Task difficulty
  - Low inter-agreement
- Use distant supervision to automatically generate training instances
  - Use the existing event instances in knowledge bases
- Not so easy as relation extraction
  - Hard to label text with event instances

Automatic Data Generation

- Use the event instances in Freebase to label the plain text in Wikipedia
- However, Triggers are not given in knowledge bases
  - Find the event mentions with distant supervision assumptions
  - Detect the triggers in the texts

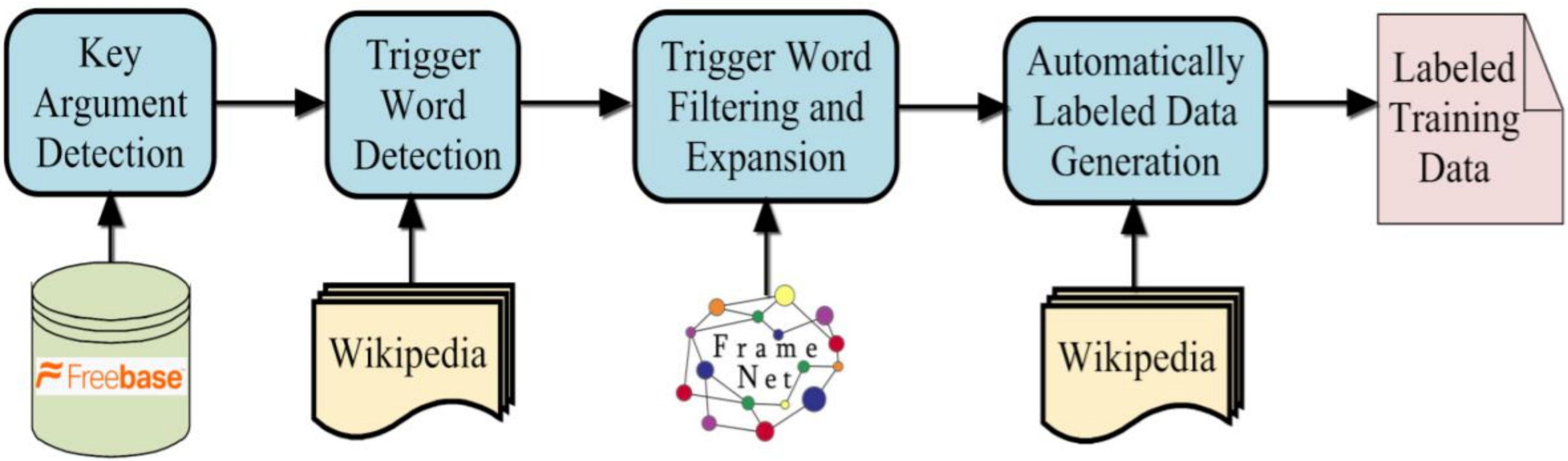
Automatically Labeled Data Generation for Large Scale Event Extraction. Chen et al. ACL 2017.

Automatic Data Generation

- Event Trigger Words Extraction
  - A trivial assumption: The sentences mention all arguments denote such events
  - Only 0.02% of instances can find all argument mentions in one sentence
  - Find the key arguments, which are usually mentioned in multiple sentences.
  - Then detect the triggers
- Argument Extraction/Role Identification
  - With given triggers and entities

Automatically Labeled Data Generation for Large Scale Event Extraction. Chen et al. ACL 2017.

Automatic Data Generation: Pipeline

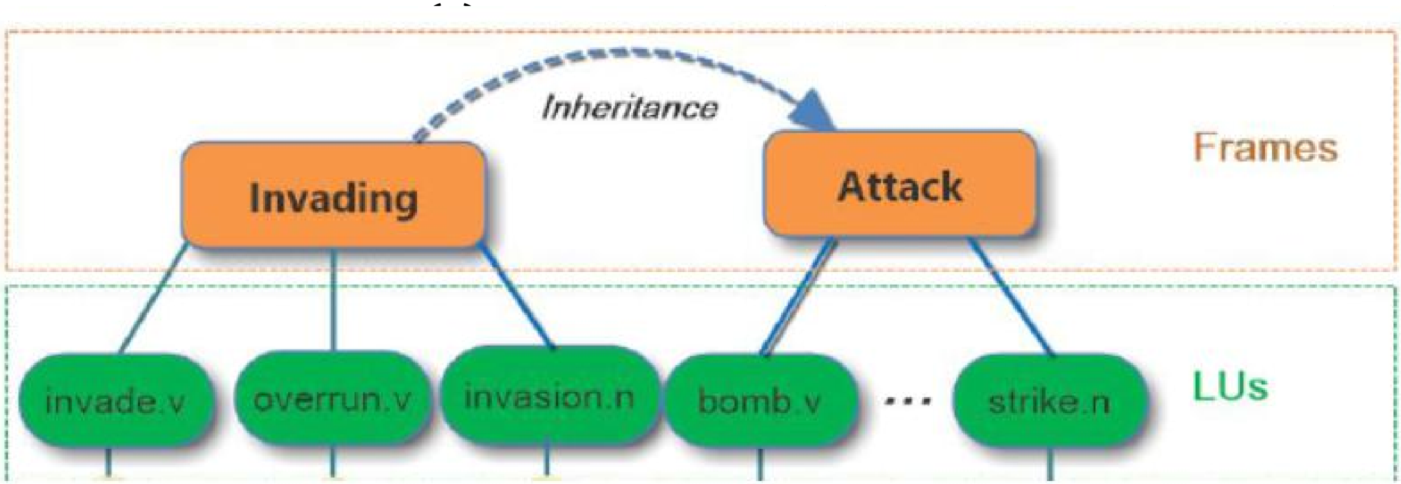


Automatic Data Generation

- Key Argument Detection:
  - Role saliency: $RS_{ij} = \frac{\text{Count}(A_i, ET_j)}{\text{Count}(ET_j)}$
  - Event Relevance: $ER_i = log \frac{\text{SUM}(ET)}{1+\text{Count}(ETC_i)}$
  - Key Rate: $KR_{ij} = RS_{ij} * ER_i$
  - Choose top K arguments with high KR as key arguments
- Trigger Word Detection:
  - Trigger Candidate Frequency： $TCF_{ij} = \frac{\text{Count}(V_i, ETS_j)}{\text{Count}(ETS_j)}$
  - Trigger Event Type Frequency： $TET\ F_i = log \frac{\text{SUM}(ET)}{1+\text{Count}(ETI_i)}$
  - Trigger Rate： $TR_{ij} = TCF_{ij} * TET\ F_i$
  - Choose verbs with high TR values as the trigger words for each event type

Automatic Data Generation

- Trigger Word Filtering and Expansion
  - Detected verbs are noisy
  - Using TR is hard to find nominal triggers
  - Use linguistic resource FrameNet to filter out noisy verbal triggers and expand nominal triggers
  - Map the event types to frames with word embedding similarity



Automatic Data Generation

- Automatically Labeled Data Generation
  - Soft Distant Supervision
  - Assume any sentence containing all key arguments in Freebase and a corresponding trigger word is likely to express that event in some way
  - Arguments occurring in that sentences are likely to play the corresponding roles in that event

Automatic Data Generation

Event of people.marriage in Freebase

id \ role	spouse	spouse	*from*	*to*	*location of ceremony*
m.02nqglv	Barack Obama	Michelle Obama	*10/3/1992*		*Trinity United Church of Christ*
...	...	...	...	...	...

Mentions from free texts

1. Michelle Obama was raised Methodist and joined the Trinity United Church of Christ, where she and Barack Obama married. ✔
2. Michelle Obama and Barack Obama married in October 1992, and have two daughters, Malia Ann and Natasha. ✔
3. Michelle Obama and Barack Obama attended the wedding of his top aids in Florida. ✖

Trigger list of people.marriage event

Automatic Data Generation

- Advantages:
  - Automatically generate relatively high-quality labeled data
  - Improve the supervised EE performance in a way
- Limitations:
  - Limited by the event instances in existing knowledge bases, which is far from sufficient
  - The generated instances are similar and not so informative
- Further reading: a bootstrapping method
  - *Adversarial Training for Weakly Supervised Event Detection. Wang et al. NAACL-HLT 2019.*

Summary

- Relation Extraction
  - Hand-written Patterns
  - Supervised RE
  - Semi-supervised RE
- Advanced Topics
  - Document-level RE
  - Open RE
  - Few-shot RE
  - Event Extraction

Reading Material

a. Relation Extraction

- Must-read papers

· Relation Classification via Convolutional Deep Neural Network. COLING 2014. [link]

· Distant Supervision for Relation Extraction without Labeled Data. ACL-IJCNLP 2009. [link]

· Neural Relation Extraction with Selective Attention over Instances. ACL 2016. [link]

- Further reading

RE paper list [link]

b. Advanced Topics

- Event Extraction

Joint Event Extraction via Structured Prediction with Global Features. ACL 2013. [link]

Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. ACL 2015. [link]

Adversarial Training for Weakly Supervised Event Detection. NAACL 2019. [link]

For reading material recommendation of this course, please refer to our github.

Reading Material

- OpenRE

Unsupervised open relation extraction. In Proceedings of European Semantic Web Conference 2017 [link]

Open Relation Extraction: Relational Knowledge Transfer from Supervised Data to Unsupervised Data. EMNLP 2019 [link]

Discrete-state variational autoencoders for joint discovery and factorization of relations. TACL 2016 [link]

- Document-Level RE

DocRED: A Large-Scale Document-Level Relation Extraction Dataset. ACL 2019 [link]

A Walk-based Model on Entity Graphs for Relation Extraction. ACL 2017 [link]

Graph Neural Networks with Generated Parameters for Relation Extraction. ACL 2019 [link]

- Few-shot RE

FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. ACL 2019 [link]

Matching Networks for One Shot Learning. [link]

Prototypical Networks for Few-shot Learning [link]

For reading material recommendation of this course, please refer to our github.

THUNLP