Are noisy sentences useless for distant supervised relation extraction?

Yu-Ming Shang, He-Yan Huang, Xian-Ling Mao, Xin Sun, Wei Wei

1. Problem Statement

- Noisy labeling problem has been one of the major drawbacks for relation extraction task.
    - relation extraction?
- Are noisy sentences truly useless?
    - Not caused by a lack of useful information, but the **missing credible relation labels**
- How do we solve this?
    - By implementing unsupervised deep clustering to generate reliable labels for noisy sentences

Relation Extraction

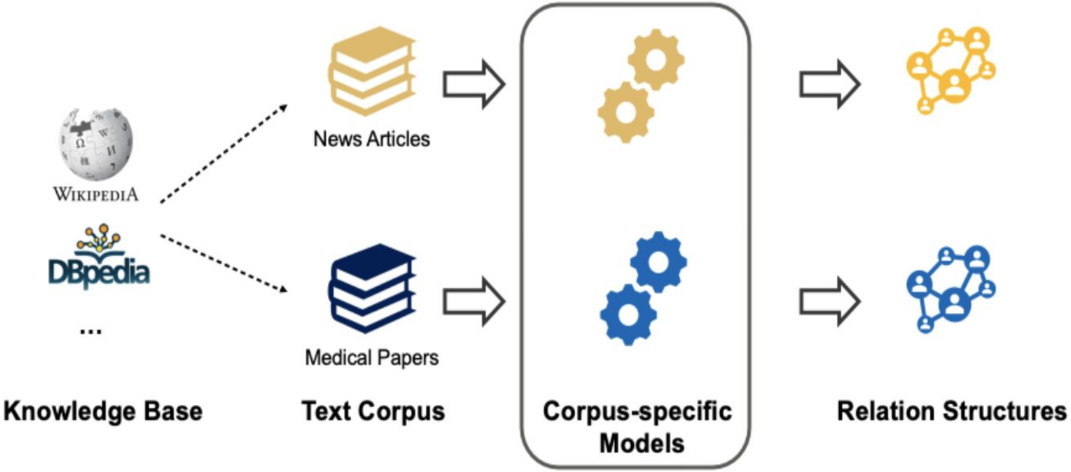
- Relation Extraction (RE) is *the task of extracting semantic relationships from text*, which usually occur between two or more entities.
- The task can be done via rule-based/weakly supervised/distantly supervised/unsupervised learning.

Sentence	Relation
1. **Steve Jobs** and Wozniak co-founded **Apple** in 1976.	*Founder*
2. **Michael Jordan** is an American retired professional **basketball player**.	*Career*
3. **Washington D.C.** is the capital of **United states**.	*CapitalOf*
......	......

Distant Supervision

It utilizes an existing Knowledge Base (KB), such as Wikipedia, DBpedia, Wikidata, Freebase, Yago, to automatically construct training data.

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care



Distant supervision

Assumption : if two entities (e1, e2) have a relationship r in knowledge graph, then any sentence that mentions the two entities might express the relation r

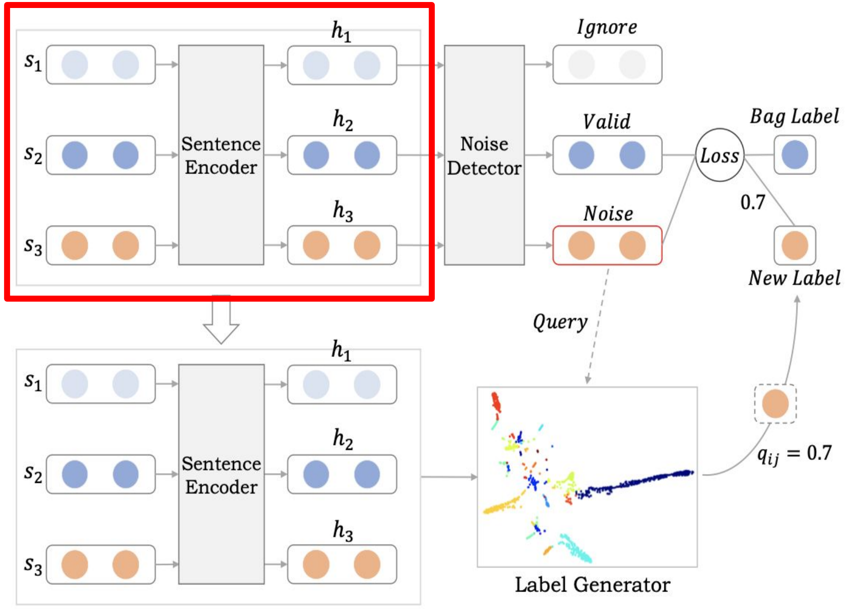
	Sentence	Bag Label	Noise?	Correct Label
Bag	#1: **Barack Obama** was born in the **United States**.	president of	Yes	born in
	#2: **Barack Obama** was the first African American to be elected to the president of the **United States**.		No	president of
	#3: **Barack Obama** served as the 44th president of the **United States** from 2009 to 2017.		No	president of

2. Method

- The paper proposes a Deep Clustering based Relation Extraction model (DCRE) that could generate reliable labels for nosity sentences.
- DCRE consists of three Modules : a sentence encoder, a noise detector and a label generator.
- Perks of a DCRE model?
    - The model can convert the noisy sentences into meaningful training data, which also leads to the increase of the number of useful sentences

2. Method : a sentence encoder

1. Transform sentences into low-dimensional vectors with word embeddings and position embeddings
    a. Position embeddings : make the model pay more attention to the words close to the target entities by calculating a series of relative distances from the current word to the two entities
2. Employ PCNN as a feature extractor
    a. each feature map Mi is divided into three parts { Mi1, Mi2, Mi3 } by the position of two entities. Then, the max-pooling operation is performed on the three parts separately.

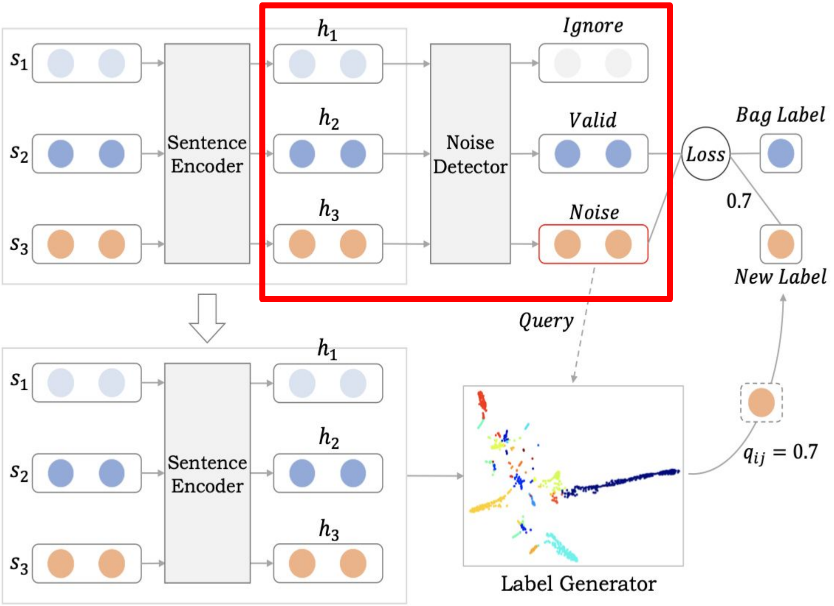


2. Method : a noise detector

1. Calculates a coefficient value with a simple dot product between the sentence representation and relation label matrix

$$a_i = \boldsymbol{h}_i \boldsymbol{l}_j^T .$$

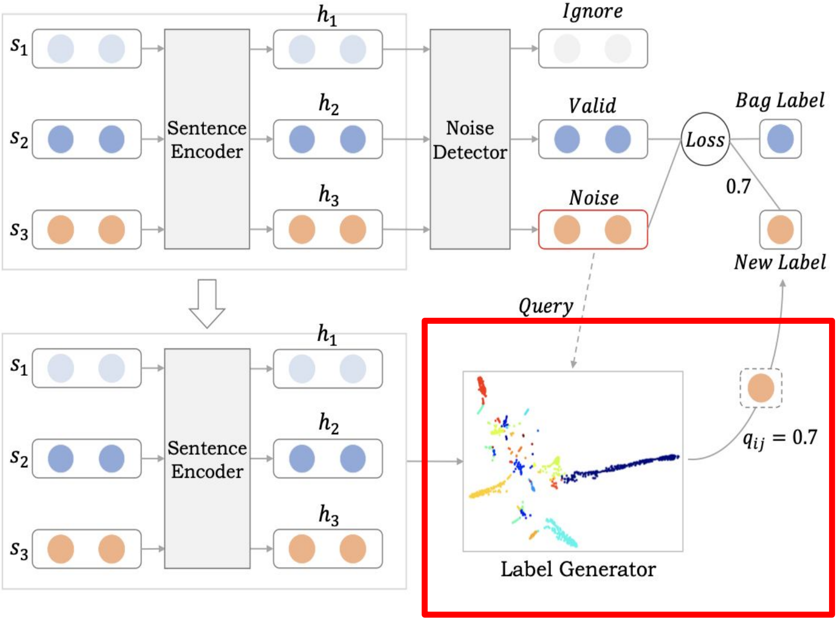
2. If the coefficient is smaller than a threshold -> noisy/ The sentence with best coefficient score -> valid / The remaining -> ignore



2. Method : a label generator

1. Employs an unsupervised deep clustering and measures the similarity between the feature vector and cluster centers via t-distribution
2. Implements a threshold for validation and introduces a scaling factor, the calculated similarity measures, as weight to scale the cross-entropy loss function

$$\mathcal{J}(\theta) = -\sum_{(x_i, y_i) \in \mathbb{V}} logp(y_i | x_i; \Theta) - \lambda \sum_{(x_i, y_i) \in \mathbb{N}} q_{ij} logp(y_j | x_i; \Theta),$$

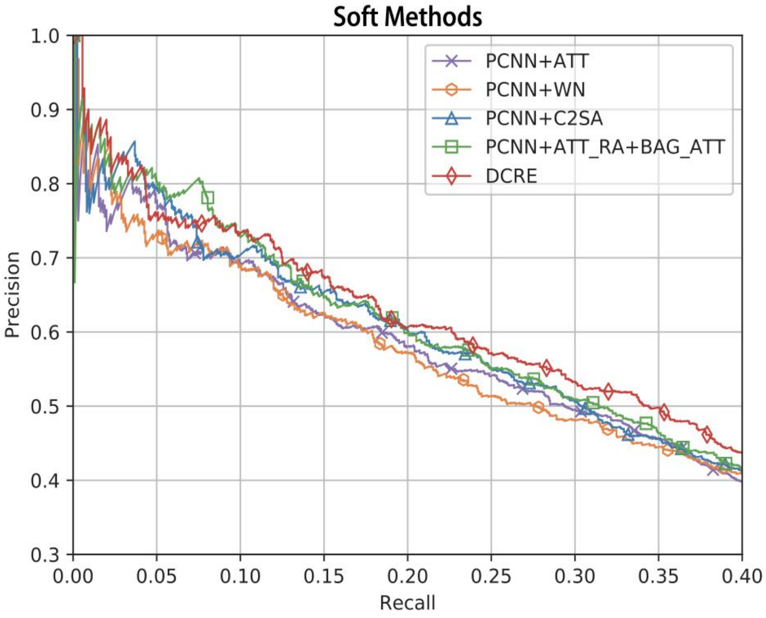


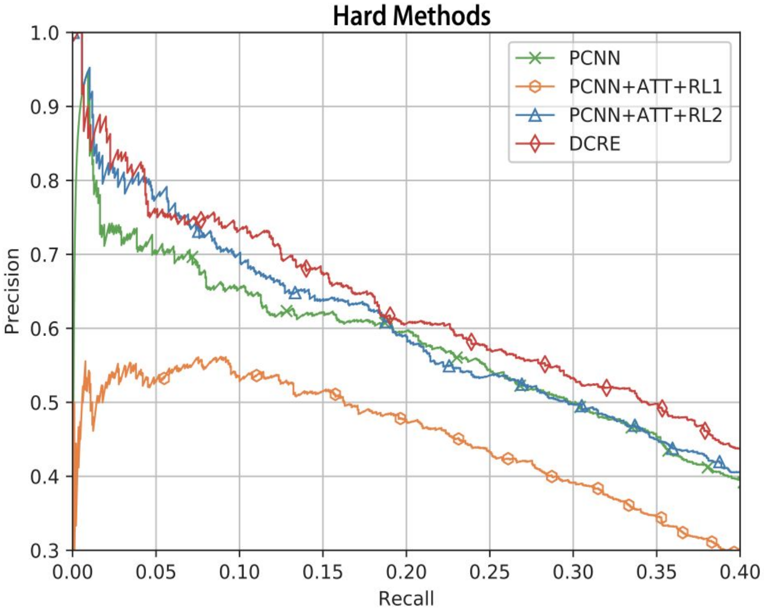
3. Experiments

- Dataset: NYT-10 which was constructed by aligning relation facts in Freebase with the New York Times corpus
  - it contains 522,611 sentences, 281,270 entity pairs in the training data; and 172,448 sentences, 96,678 entity pairs in the test data ; 53 relations in total
- employed *k*-means for clustering, obtain multiple clustering results and determine its final category by voting
- For evaluation, the relations extracted from testing data are automatically compared with those in Freebase
- Compared the performance with 7 different baseline models with precision-recall curves

4. Results

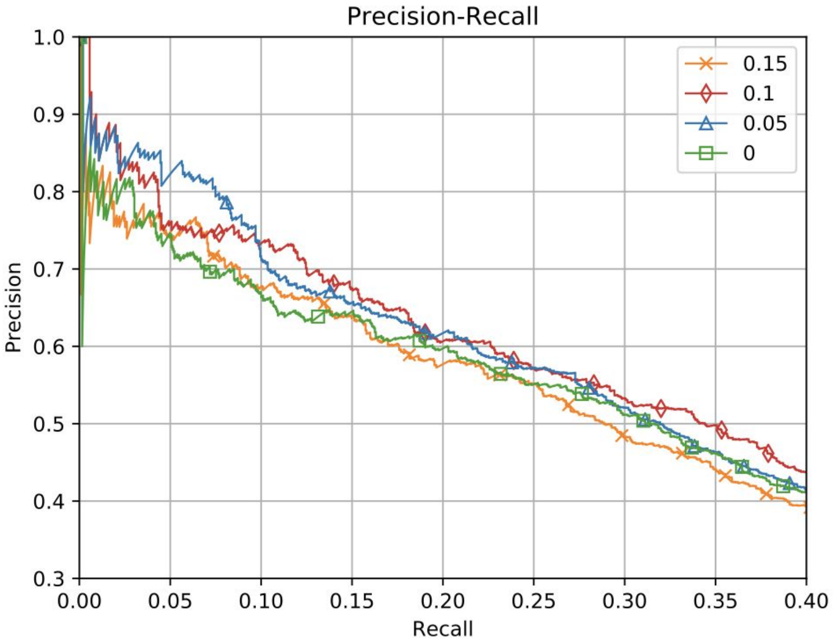
Soft methods (place soft weights on sentences to reduce the impact of noisy sentences) vs Hard methods (removes all the noisy sentences)





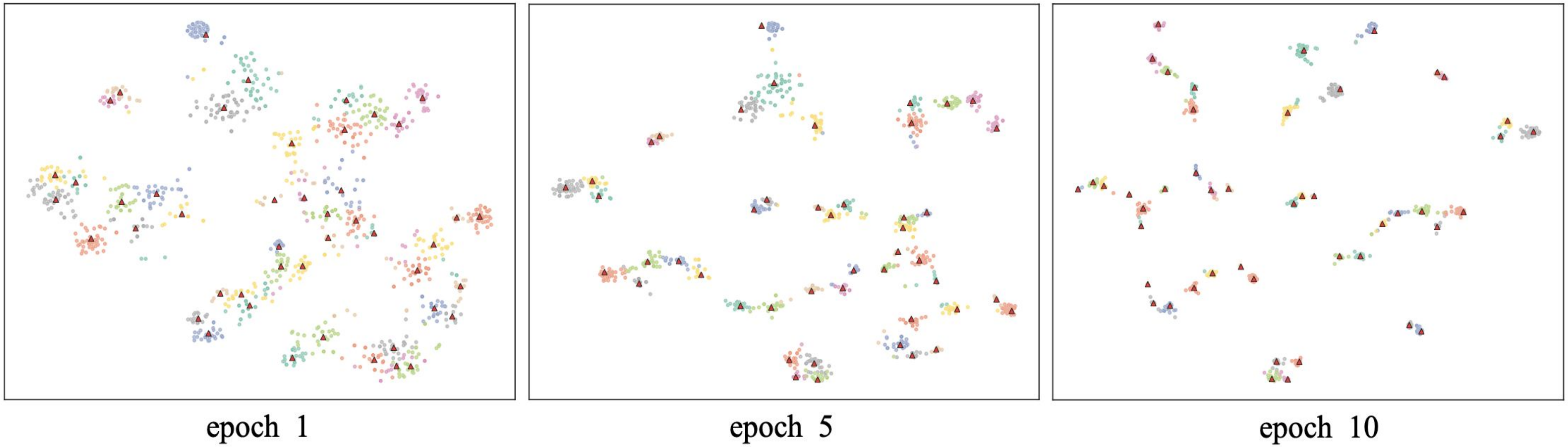
4. Results

- Manually tested accuracy of threshold value ranging from {0.15, 0.1, 0.05, 0}
- 0.1 demonstrates the best performance



4. results

set the number of clusters as 47, excluding 6 long-tail relations which appear less than 2 times in training data



4. Results

Correct label for a 1st pair is */location/country/capital* and a 4th pair is *people/person/place lived*

ID	Entity pair	Sentence	Original label	Generated label	Correct?
1	(China,Beijing)	**Beijing** has tried to enlist the support of Uzbekistan in fighting Islamic separatism in **China**'s western region of Xinjiang, while also lining up secure supplies of oil and gas.	/location/location/contains	/location/cn province /capital	No
2	(Italy, Rome)	Mr. Tomassetti's companies are named after L'Aquila, **Italy**, his birthplace 58 miles northeast of **Rome**.	/location/country/capital	/location/location/contains	Yes
3	(Saddam Hussein, Iraq)	As national journal reported in April, it was Senator Roberts who stated as the **Iraq** war began that the U.S. had "human intelligence that indicated the location of **Saddam Hussein**."	/people/deceased person/place of death	/people/person/place lived	Yes
4	(Edith Sitwell, England)	His first book was published privately in his own country and then by a major publisher in **England**, where he had many supporters in the literary world, most notably **Edith Sitwell** and Angus Wilson.	/people/person/nationality	/people/person/place of birth	No
5	(Louisiana, New Orleans)	The book, by a **New Orleans** resident, John M. Barry, describes the history and politics behind a flood that killed 1,000 people and displaced 900,000 from **Louisiana** to Illinois.	/location/location/contains	NA	Yes

5. Future work

- multi-class clustering
- automated noisy sentence selection

Thanks