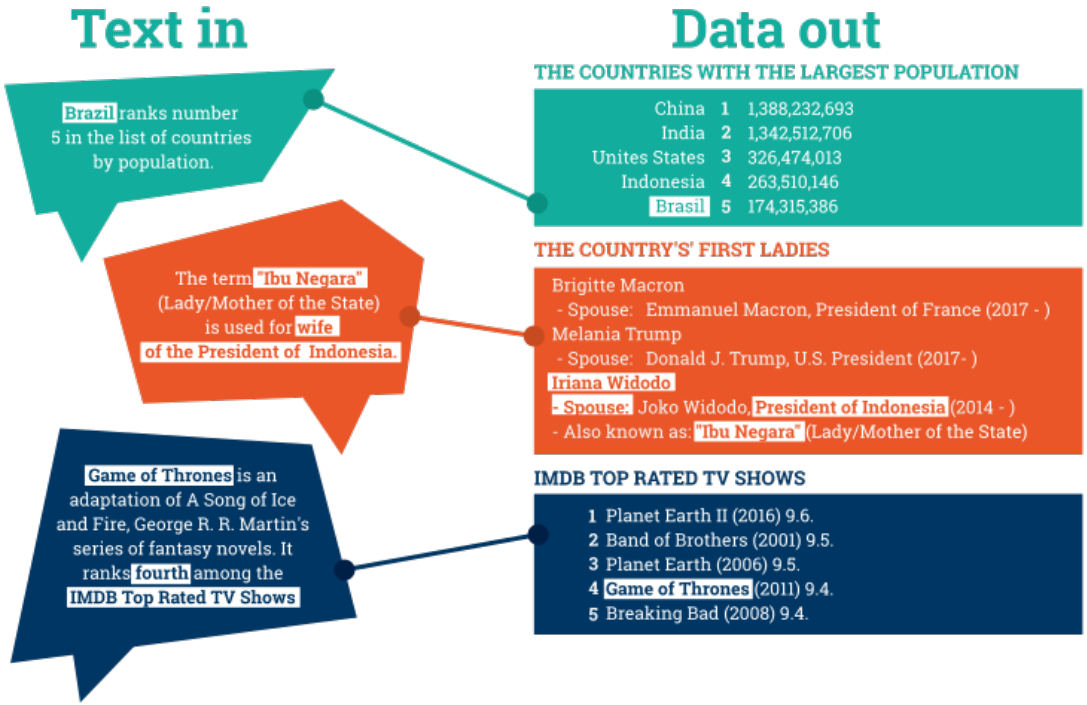
Towards Neural Information Extraction without Manual Annotated Data

Peng Xu

Department of Computing Science
University of Alberta

Table of Contents

Information Extraction



The process of Information Extraction (IE) is the task of automatically extracting structured information from unstructured documents.

Example

When you read one news article started with:

Steve Kerr has turned down **Phil Jackson** and the **New York Knicks** to accept a five-year, $25 million offer to become the **Golden State Warriors**' next coach, saying "it just felt like the right move on many levels."

To understand the information encoded in this sentence, you will need to at least know the people and organizations mentioned and the semantic relations among them.

Tasks

- The task of **named entity recognition** (NER) is to find each mention of a named entity in the text and label its type.
- As compared to traditional NER, **Fine-grained entity type classification** (FETC) works on a much larger set of fine-grained types which form a tree-structured hierarchy.
- The task of **relation extraction** (RE) is to find and classify semantic relations among entity mentions.

IE without Manual Annotated Data

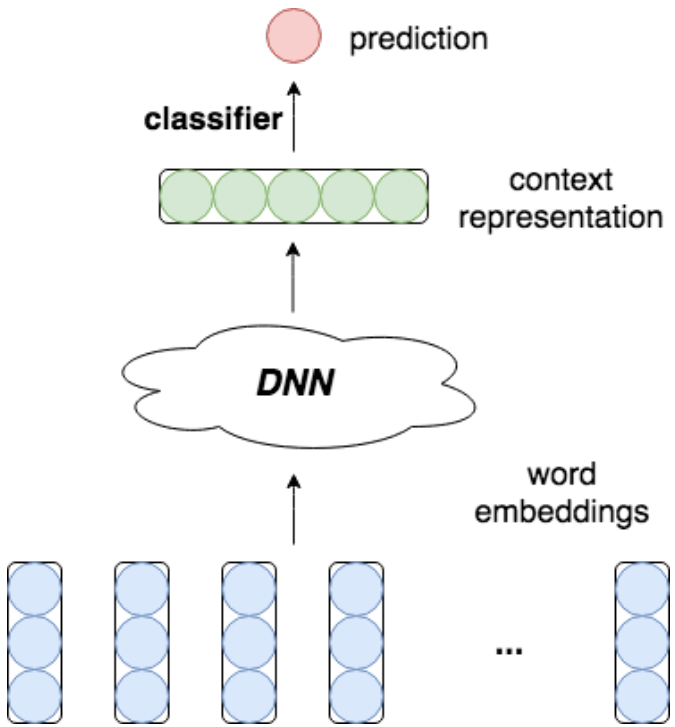
- Challenge: the absence of human-annotated data
- Approach: resort to distant supervision and annotate training corpora automatically using KBs
- Disadvantage: introduction of label noise

A typical workflow of distant supervision:

1. identify entity mentions in the documents
2. link mentions to entities in KB
3. assign, to each entity mention or entity pair, all types or relations associated in a KB

Deep Neural Networks for NLP

Over the past few years, we have witnessed the huge success of neural networks as powerful machine-learning models, yielding state-of-the-art results in many fields including NLP. As a result, we use DNNs as the backbone of our proposed models.



Knowledge Base Embedding

- Knowledge bases (KBs) enable various real-world applications.
- A major challenge to use KBs: lack of capability of accessing the similarities among different entities and relations.
- The main idea of knowledge base embedding (KBE) techniques is to represent the entities and relations in a vector space.
- These embeddings can help IE tasks by incorporating knowledge information.

Table of Contents

The Task: Fine-Grained Entity Type Classification

- Traditional *Coarse-Grained Entity Type Classification*, as a sub-task of *Named Entity Recognition (NER)*, focuses on a small set of coarse types.
- *Fine-Grained Entity Type Classification (FETC)* aims at labeling entity mentions in context with one or more specific types organized in a hierarchy.

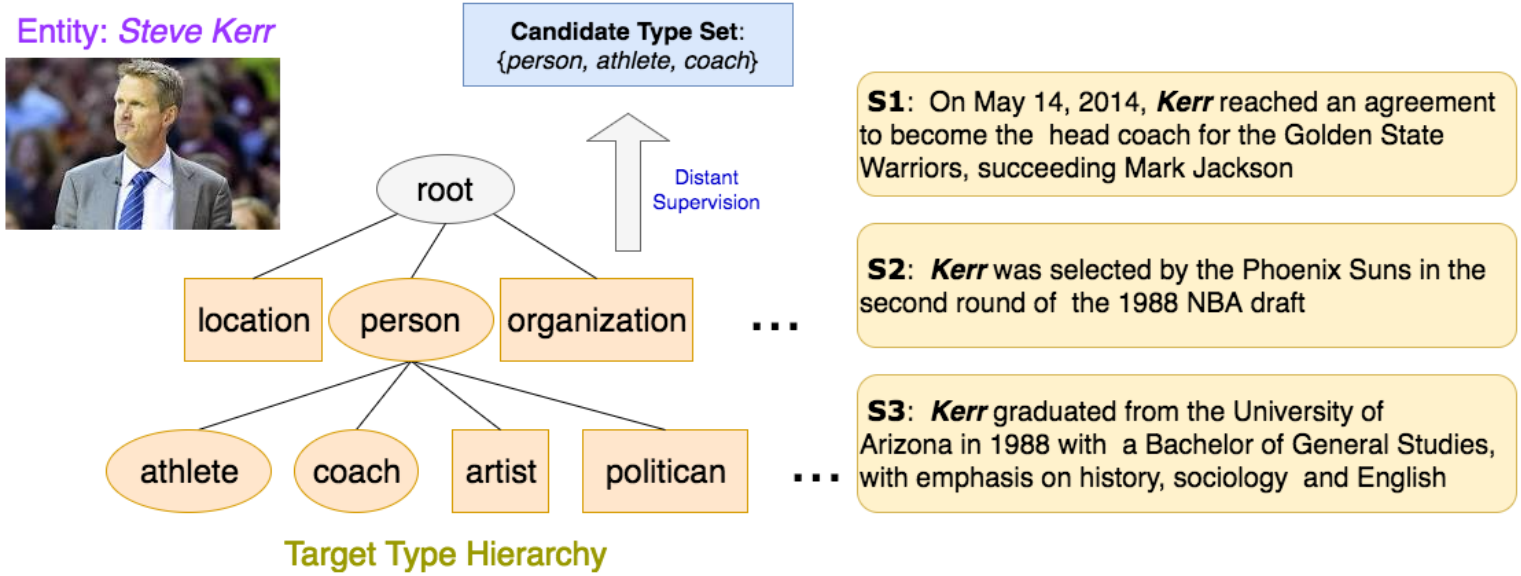
Figure: Traditional coarse-grained types are colored in black. Fine-grained types are colored in red.

Motivation

Fine-grained types help in many applications:

- relation extraction
- question answering
- coreference resolution
- entity linking
- knowledge base completion
- entity recommendation
- and so on...

Characteristics of FETC



- Context dependent labeling
- Hierarchical structure of entity types
- Collapse of the mutual exclusion assumption
- Noise in automatically annotated data

Multi-label vs. Single label

In FETC, types are not mutually exclusive!
It is natural to formulate the task as a multi-label classification problem and most FETC methods adopt this setting.
However,

- context dependent labeling $\rightarrow$ assumption that one mention can only have one *type-path* along the hierarchy
- type hierarchy is a tree $\rightarrow$ each *type-path* can be uniquely represented by the terminal type (not necessarily a leaf node)

Then, the task can be transformed to predict the terminal type of the *type-path* in the hierarchy, which is a single-label classification problem!

Pros and Cons to Adopt Single Label Setting

Pros:

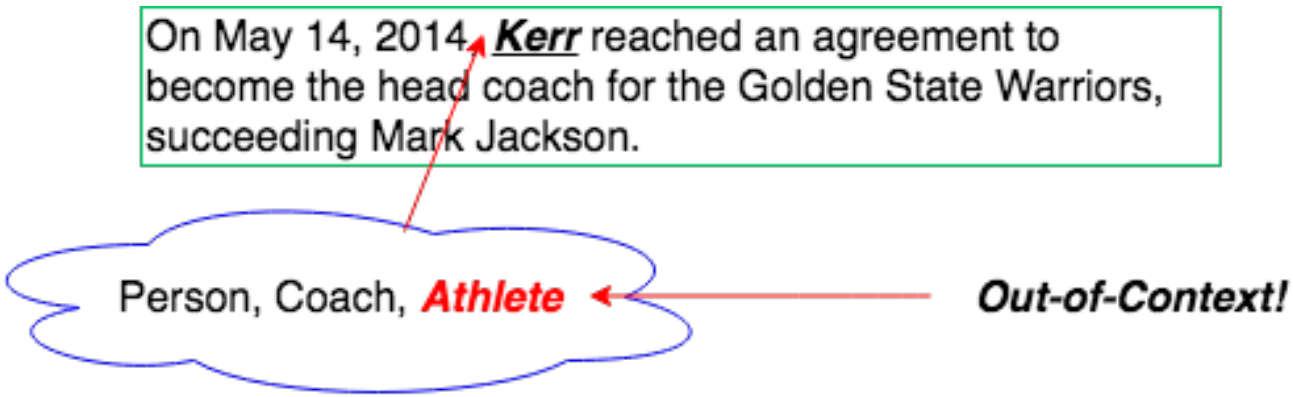
1. Simpler and more elegant
2. Benefit from previous research
3. No post-processing needed

Cons:

1. The upper bounds are no longer 100% (But, is that really important? State-of-the-art methods are nowhere near 80% strict accuracy.)

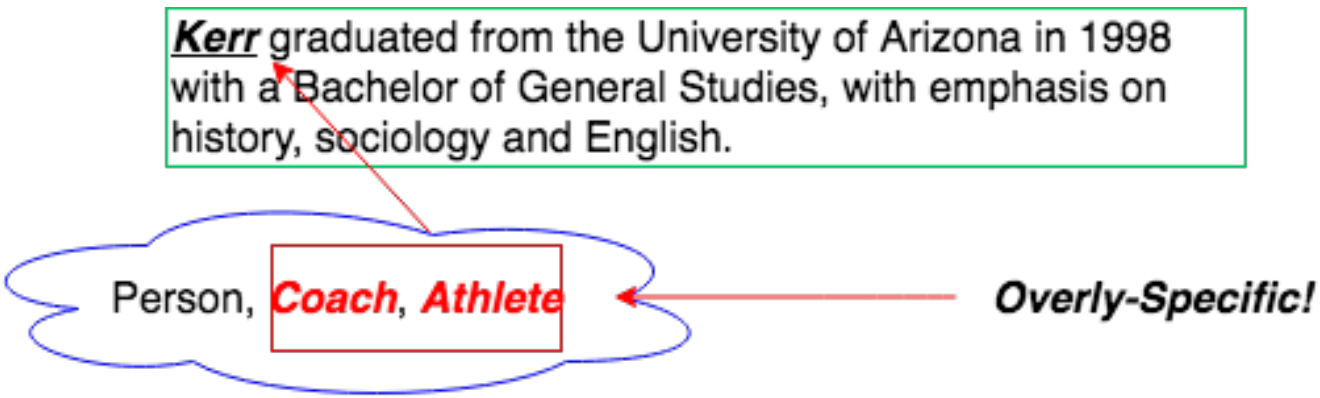
	FIGER(GOLD)	OntoNotes
# types	113	89
# raw testing mentions	563	8963
% testing mentions with single *type-path*	88.28	94.00

Out-of-context Noise



One kind of noise introduced by distant supervision is assigning labels that are *out-of-context*.

Overly-specific Noise



Another source of noise introduced by distant supervision is when the type is *overly-specifc* for the context.

Typical FETC Methods

	Attentive	AFET	LNR	A&A
without manual features	✗	✗	✗	✓
use attentive neural network	✓	✗	✗	✗
adopt single label setting	✗	✗	✗	✗
handle *out-of-context* noise	✗	✓	✓	✓
handle *overly-specifc* noise	✗	✓	✓	✗

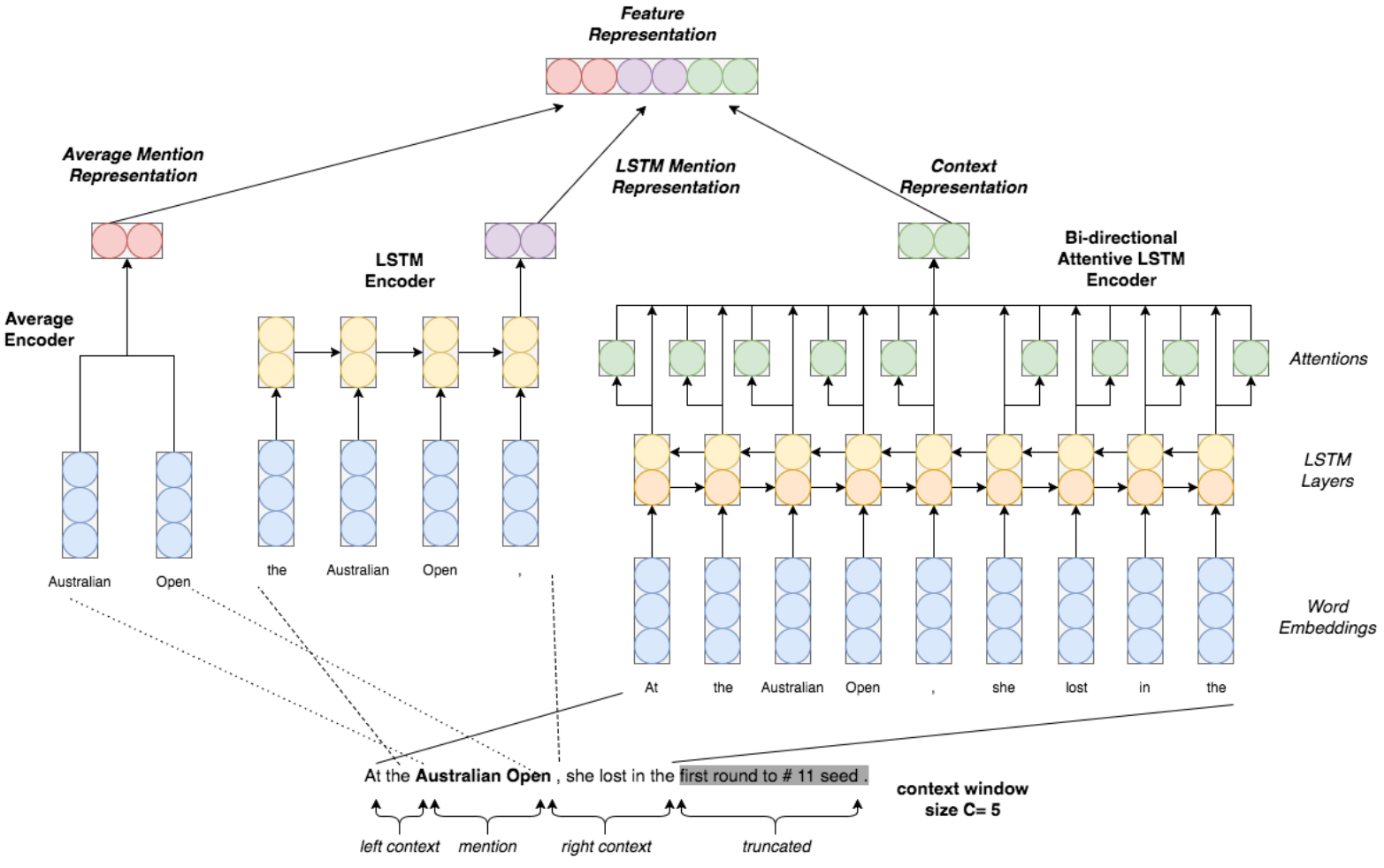
1. Attentive: Shimaoka *et al.* (2017)
2. AFET: Ren *et al.* (2016a)
3. LNR: Ren *et al.* (2016b)
4. A&A: Abhishek and Awekar (2017)

Our Proposed Model: NFETC

NFETC is a single, much simpler and more elegant neural model that attempts FETC "end-to-end" without post-processing or ad-hoc features.

	Attentive	AFET	LNR	A&A	NFETC
without manual features	✗	✗	✗	✓	✓
use attentive neural network	✓	✗	✗	✗	✓
adopt single label setting	✗	✗	✗	✗	✓
handle *out-of-context* noise	✗	✓	✓	✓	✓
handle *overly-specifc* noise	✗	✓	✓	✗	✓

Neural Architecture



A Simple Yet Effective Variant of Cross-Entropy

Traditional cross-entropy loss:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log(\hat{p}(y_i)), \tag{1}$$

which can't handle data with multi *type-paths* (that is, with *out-of-context* noise). A simple yet effective variant of the cross-entropy loss:

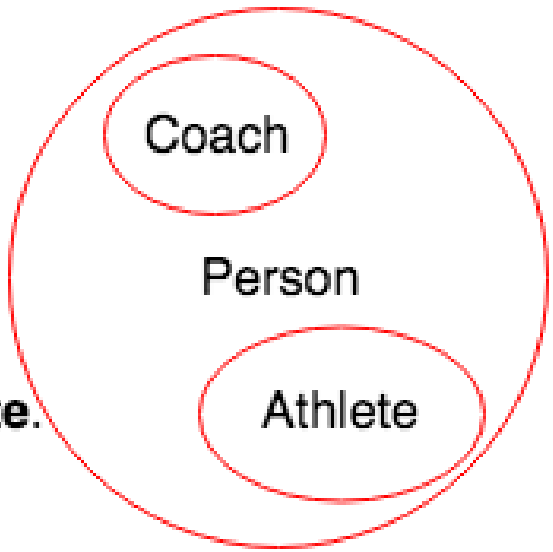
$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log(\hat{p}(y_i^*)), \tag{2}$$

where $y_i^* = \arg\max_{y \in \mathcal{Y}_i} \hat{p}(y)$ and $\mathcal{Y}_i$ is the labelled type set.

Hierarchical Loss Normalization: Intuition

On May 14, 2014, ***Kerr*** reached an agreement to become the head coach for the Golden State Warriors, succeeding Mark Jackson.

Person, Coach



What if we predict ***Kerr*** as **Person** here?
It's correct in some sense compared to **Athlete**.
Types are correlated!

Hierarchical Loss Normalization [1]

Based on the intuition, we adjust the estimated probability to:

$$p^*(\hat{y}) = p(\hat{y}) + \beta * \sum_{t \in \Gamma} p(t) \tag{3}$$

where Γ is the set of ancestor types along the *type-path* of $\hat{y}$, β is a hyperparameter. In this way, the model will:

1. get less penalty when it predicts the actual type for data with *overly-specific* noise
2. prefer generic types unless there is a strong indicator for a more specific type in the context

[1] Hierarchical loss function (Cai and Hofmann, 2004) was originally introduced in the context of document categorization with SVM. However, they assume that weights to control the hierarchical loss can be solicited from domain experts which is inapplicable for FETC.

Experiments

- Datasets: FIGER(GOLD) & OntoNotes

	FIGER(GOLD)	OntoNotes
# types	113	89
# raw training mentions	2009898	253241
# raw testing mentions	563	8963
% filtered training mentions	64.46	73.13
% filtered testing mentions	88.28	94.00
Max hierarchy depth	2	3

- Evaluation Metrics: Strict Accuracy, Macro F1 and Micro F1

Results

Model	FIGER(GOLD)			OntoNotes		
	Strict Acc.	Macro F1	Micro F1	Strict Acc.	Macro F1	Micro F1
Attentive	59.68	78.97	75.36	51.74	70.98	64.91
AFET	53.3	69.3	66.4	55.1	71.1	64.7
LNR+FIGER	59.9	76.3	74.9	57.2	71.5	66.1
A&A	65.8	**81.2**	77.4	52.2	68.5	63.3
NFETC(f)	57.9 $\pm$ 1.3	78.4 $\pm$ 0.8	75.0 $\pm$ 0.7	54.4 $\pm$ 0.3	71.5 $\pm$ 0.4	64.9 $\pm$ 0.3
NFETC-hier(f)	68.0 $\pm$ 0.8	**81.4 $\pm$ 0.8**	77.9 $\pm$ 0.7	59.6 $\pm$ 0.2	**76.1 $\pm$ 0.2**	69.7 $\pm$ 0.2
NFETC(r)	56.2 $\pm$ 1.0	77.2 $\pm$ 0.9	74.3 $\pm$ 1.1	54.8 $\pm$ 0.4	71.8 $\pm$ 0.4	65.0 $\pm$ 0.4
NFETC-hier(r)	**68.9 $\pm$ 0.6**	**81.9 $\pm$ 0.7**	**79.0 $\pm$ 0.7**	**60.2 $\pm$ 0.2**	**76.4 $\pm$ 0.1**	**70.2 $\pm$ 0.2**

Variants of our proposed model:

- NFETC(f): basic model trained on data with single *type-path*
- NFETC-hier(f): with hierarchical loss normalization trained on data with single *type-path*
- NFETC(r): with variant of cross-entropy trained on raw data
- NFETC-hier(f): with variant of cross-entropy and hierarchical loss normalization trained on raw data

Case Study

Test Sentence	Ground Truth	Prediction (w/o HLN)	Prediction (w HLN)
S1: **Hopkins** said four fellow elections is curious , considering the . . .	**Person**	**Politician**	**Person**
S2: . . . for WiFi communications across all **the SD cards.**	**Product**	**Software**	**Product**

Test Sentence	Ground Truth	Prediction (original CE)	Prediction (improved CE)
S3: **ASC** Director Melvin Taing said that because the commission is . . .	**Organization**	**Title**	**Organization**

Test Sentence	Ground Truth	Failed Prediction
S4: A handful of professors in the **UW** Department of Chemistry . . .	**Educational Institution**	**Organization**
S5: Work needs to be done and, in **Washington state**, . . .	**Province**	**City**

Table of Contents

Task and Motivation

- Knowledge Bases (KBs) are used in support of many important NLP applications.
- Building KBs is a non-trivial and never-ending task.
- As the world changes, new knowledge needs to be harvested while old knowledge needs to be revised.
- Relation Extraction: assign a KB relation to a sentence containing a pair of entities, which in turn can be used for updating the KB.

Wrong Labelling Problem

Although distant supervision is an effective strategy to automatically label training data, it always suffers from wrong labelling problem.

Example 1: Relation *founder_of*

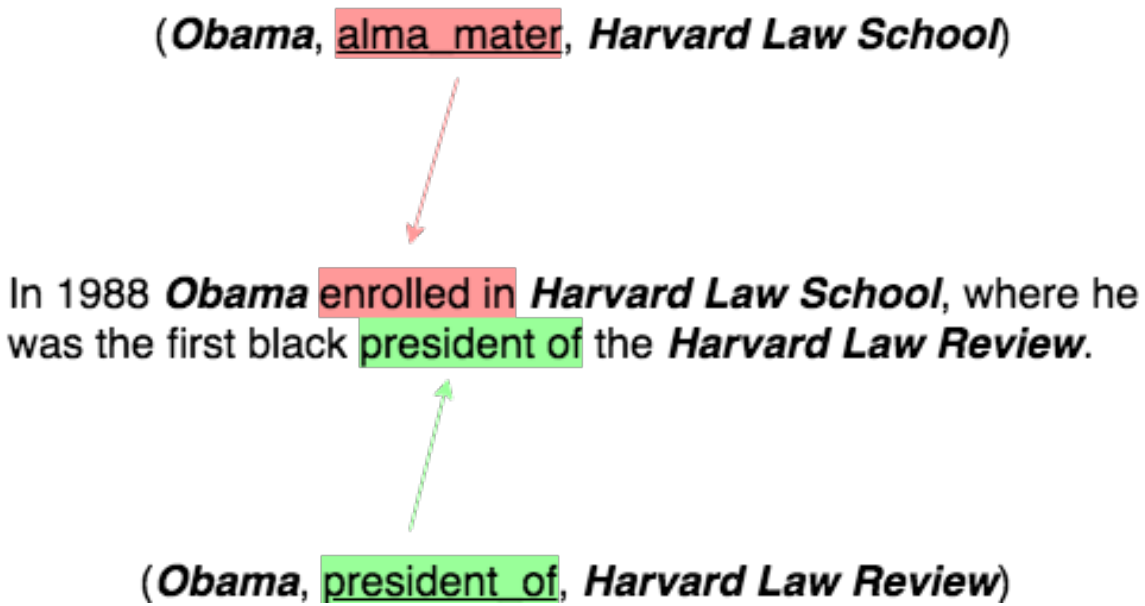
Correct: Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.
Wrong: Steve Jobs passed away the day before Apple unveiled iPhone 4S in late 2011.

Example 2: Relation *president_of*

Correct: Barack Obama is an American politician who served as the 44th President of the United States.
Wrong: Barack Obama was born in 1961 in the United States.

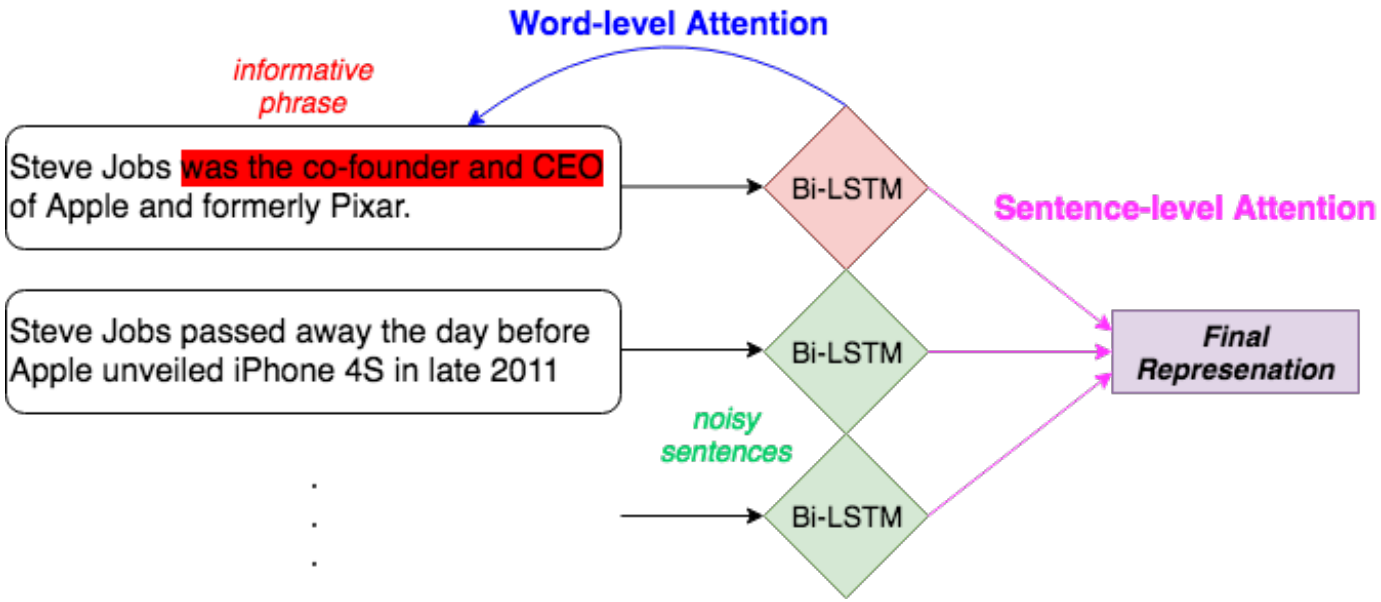
Attention Mechanisms

- Human visual attention is able to focus on a certain region of an image with "high resolution" while perceiving the surrounding image in "low resolution".
- For NLP, it can help the model distinguish which parts of the given texts are more indicative for the task.



Bi-LSTM with Multi-Level Attention Mechanisms

- Bi-LSTM as the backbone of our model.
- Word-level attention to capture the most informative phrase.
- Sentence-level attention to address the wrong labelling problem.



Experiments

- Dataset: NYT
  - align Freebase relations mentioned in the New York Times Corpus
  - Articles from years 2005-2006 for training
  - Articles from 2007 for testing
- Evaluation Metric: Precision/Recall curves
- Baselines: Three feature-based methods and two convolutional neural network based methods
  - Mintz: Mintz *et al.* (2009)
  - MultiR: Hoffmann *et al.* (2011)
  - MIML: Surdeanu *et al.* (2012)
  - CNN+ATT: Lin *et al.* (2016)
  - PCNN+ATT: Lin *et al.* (2016)

Results

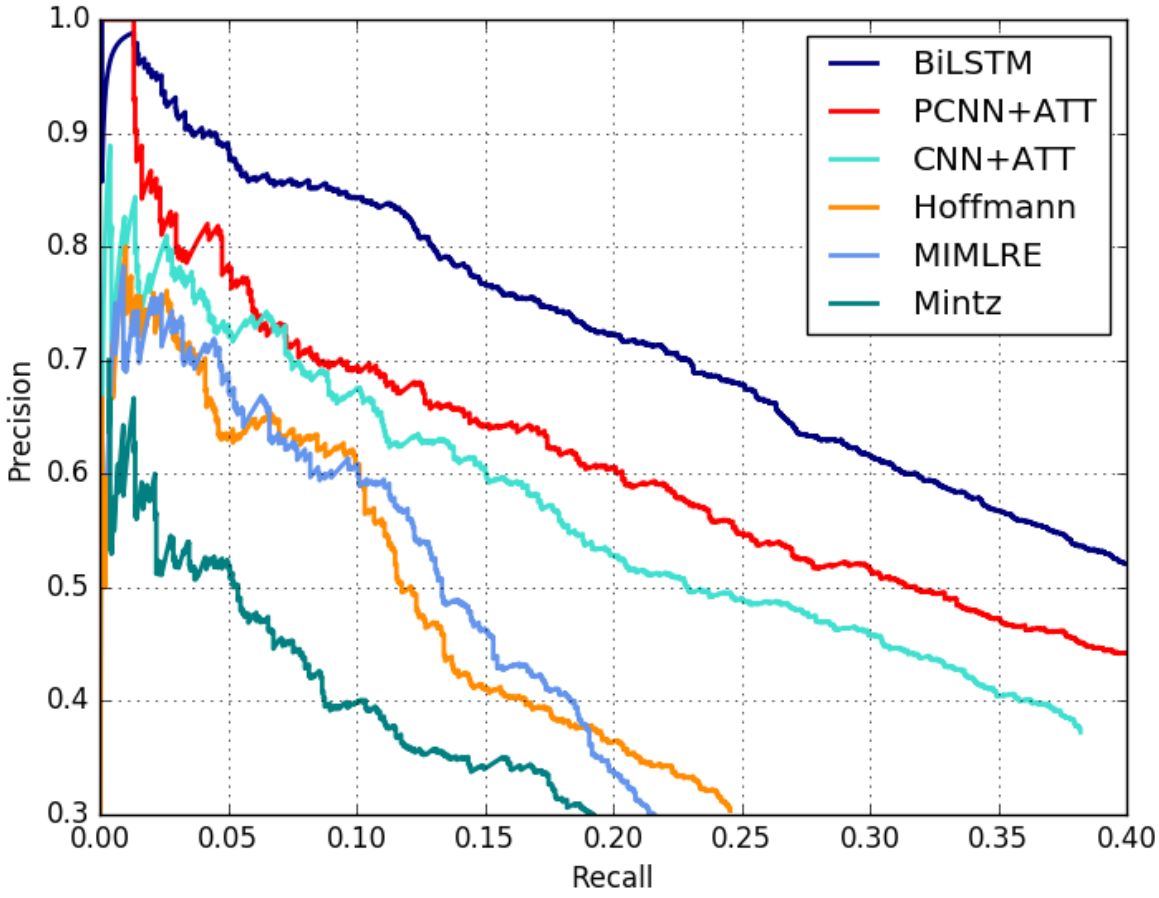

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Precision
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Recall
BiLSTM
PCNN+ATT
CNN+ATT
Hoffmann
MIMLRE
Mintz

Table of Contents

Introduction

- A related task of RE is Knowledge Base Embedding (KBE)
- Weston *et al.* (2013) was the first to show that combining predictions from RE and KBE models, trained in isolation, improves the effectiveness on the RE task.
- Their strategies are rather naive and unable to give improvements for the state-of-the-art neural RE model.

$$S_{combined} = \alpha S_{RE} + (1 - \alpha) S_{KBE}$$

Facilitating Relation Extraction with Existing Strategies

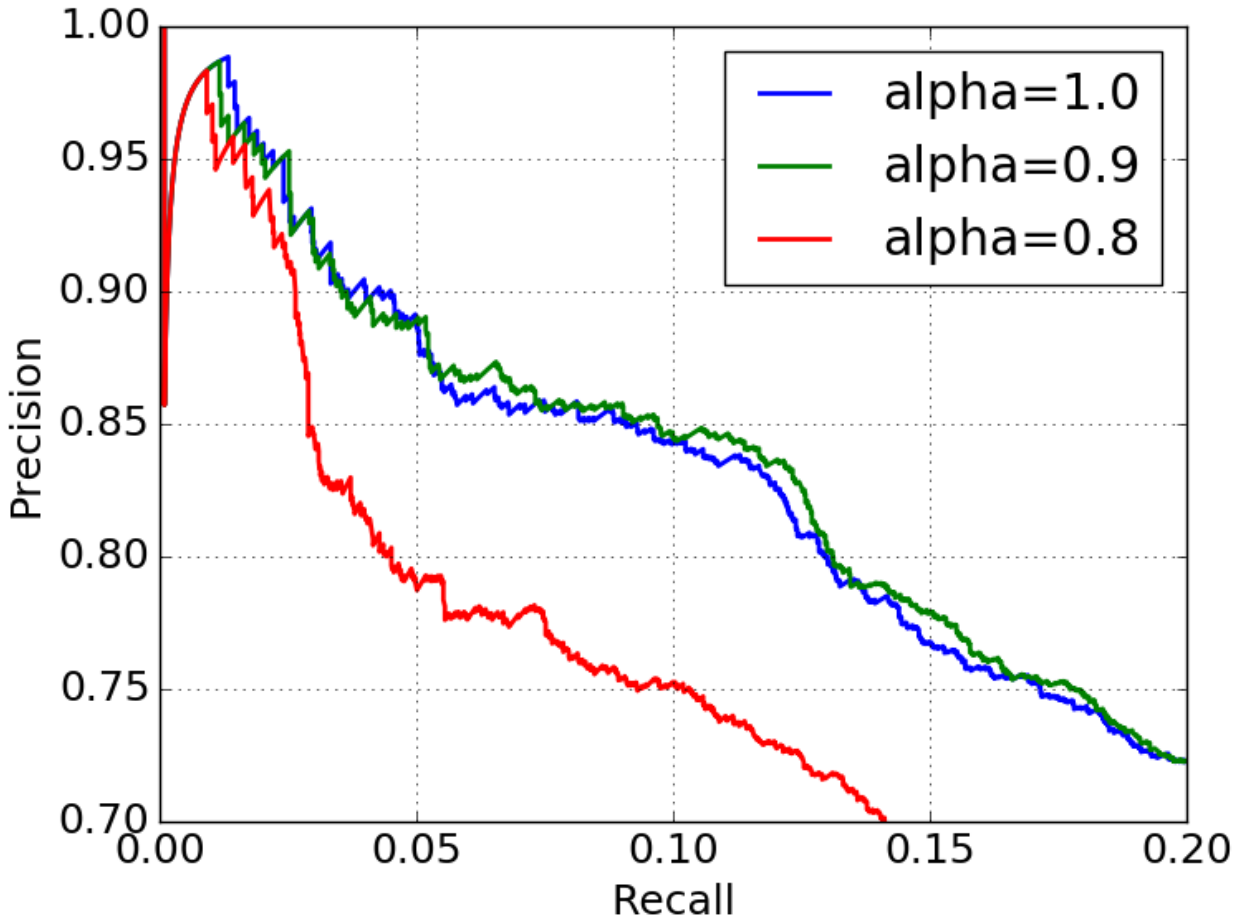


Figure: Precision-recall curves of **TransE** with different alpha.

Cont.

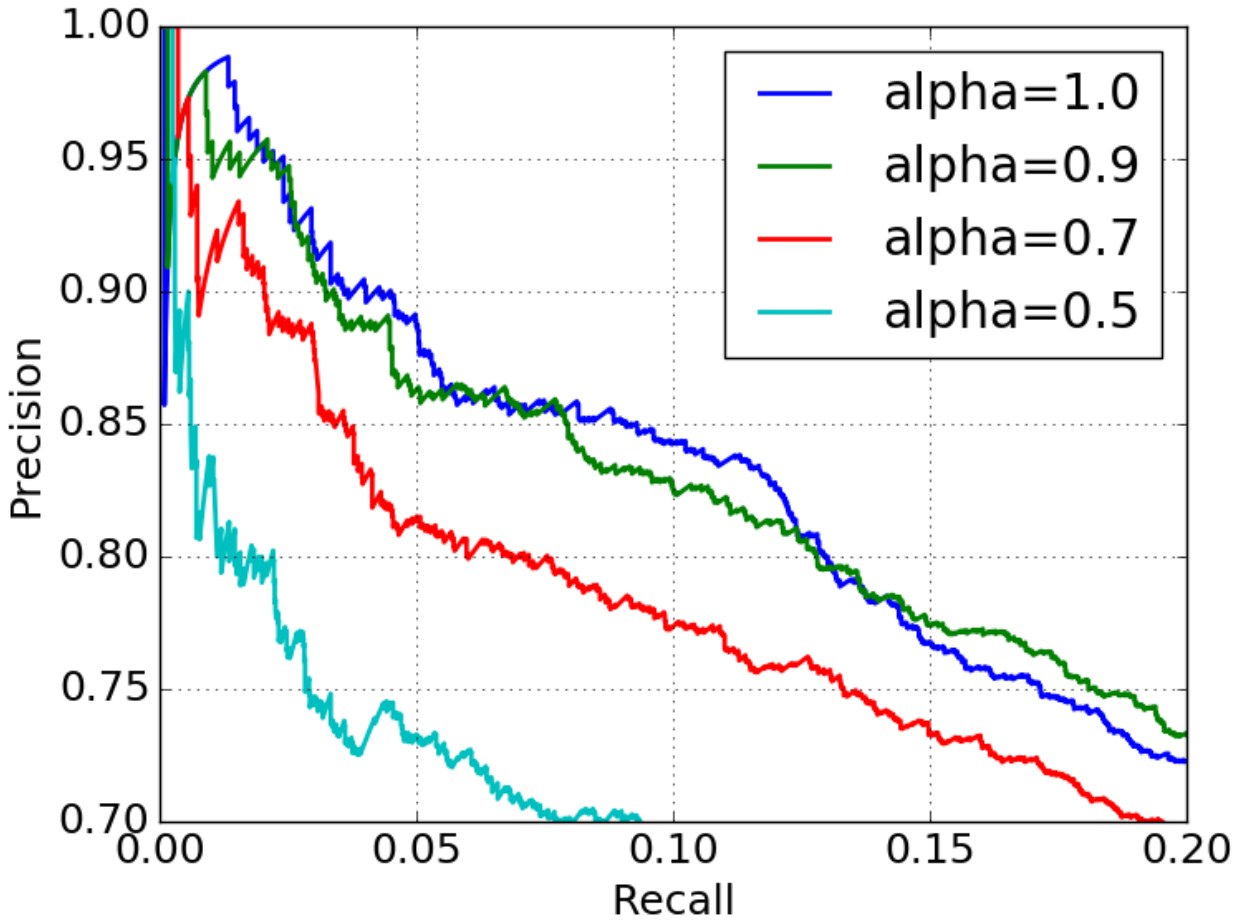
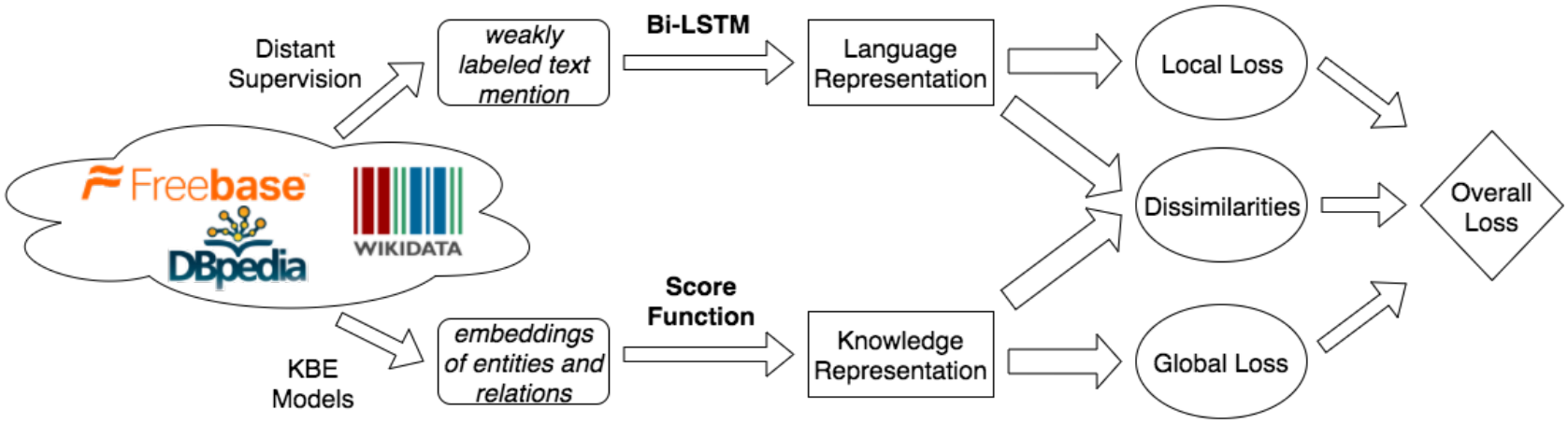


Figure: Precision-recall curves of **ComplEx** with different alpha.

Our Approach: The Overall Framework HRERE



- The language representation is learnt based on the input set of sentences $\mathcal{S}$ for each entity pair with the neural architecture described in the previous section.
- With the language representation, we can get the probability $p(r|\mathcal{S})$.

Knowledge Representation

- Following the score function ϕ and training procedure of Trouillon *et al.* (2016), we can get the knowledge representations e_h, w_r, e_t.
- With the knowledge representations and the score function, we can get the probability:

$$p(r|h,t) = \frac{e^{\phi(e_h, w_r, e_t)}}{\sum_{r'} e^{\phi(e_h, w_{r'}, e_t)}}$$

Connecting Representations

The cross-entropy losses based on the language and knowledge representations are defined as:

$$\mathcal{J}_L = -\frac{1}{N}\sum_{i=1}^{N} \log p(r_i|\mathcal{S}_i)$$

$$\mathcal{J}_G = -\frac{1}{N}\sum_{i=1}^{N} \log p(r_i|(h_i, t_i))$$

In addition, a cross-entropy loss is used to measure the dissimilarity between two distributions, thus connecting them:

$$\mathcal{J}_D = -\frac{1}{N}\sum_{i=1}^{N} p(r_i^*|\mathcal{S}_i)$$

where $r_i^* = \arg\max p(r|(h_i, t_i))$.

Model Learning

We form the joint optimization problem for model parameters as

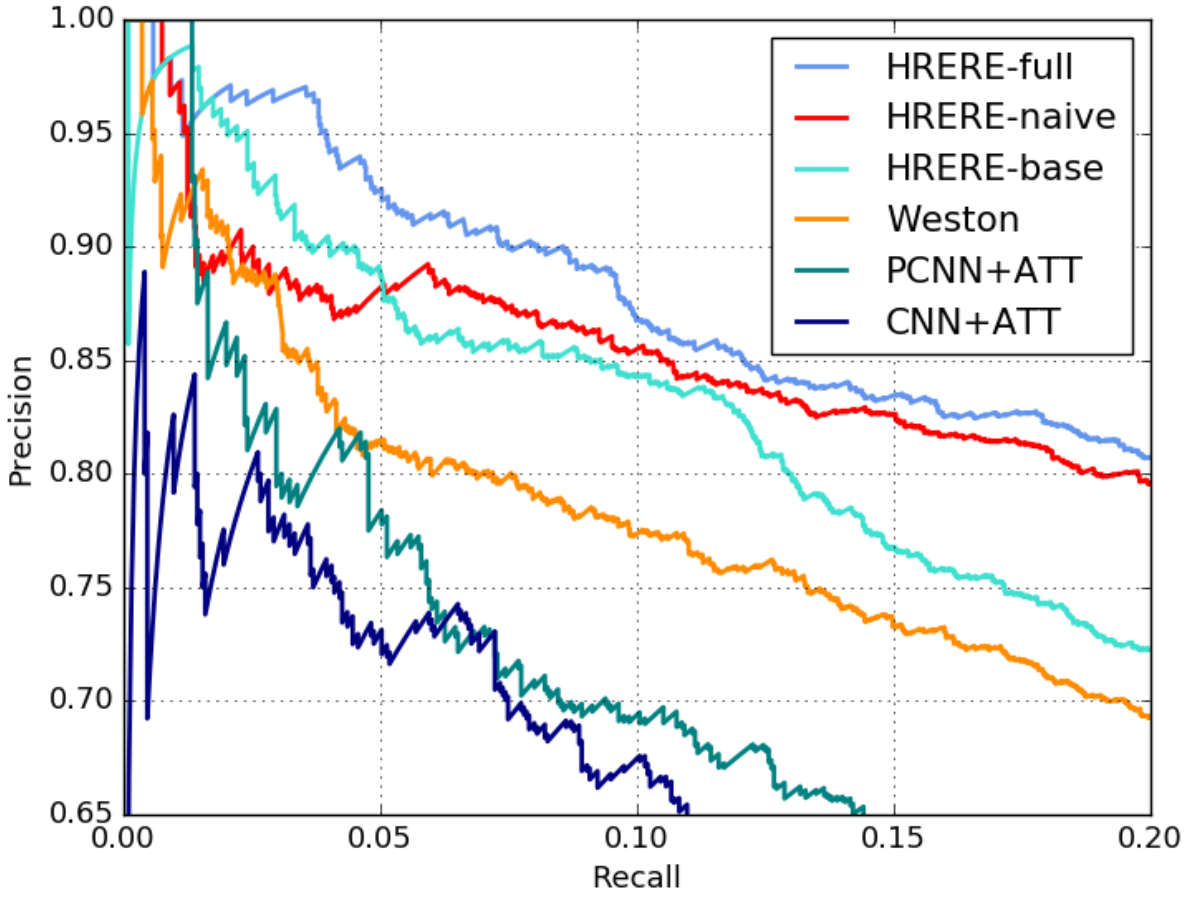
$$\min_{\Theta} \mathcal{J} = \mathcal{J}_L + \mathcal{J}_G + \mathcal{J}_D + \lambda \|\Theta\|_2^2$$

where Θ is the set of all the parameters in the model. We adopt the stochastic gradient descent with mini-batches and Adam (Kingma and Ba, 2014) to update Θ.

Experiments

- Dataset: NYT
- Knowledge Base: a Freebase subset with the 3M entities with highest degree
- Evaluation Metrics: Precision/Recall curves and P@N.
- Baselines:
  - CNN+ATT: Lin *et al.* (2016)
  - PCNN+ATT: Lin *et al.* (2016)
  - Weston: combine the scores computed with methods proposed directly without joint learning
- Variants of our proposed framework:
  - HRERE-base: basic neural model with local loss only
  - HRERE-naive: neural model with both local and global loss but without the dissimilarities
  - HRERE-full: neural model with both local and global loss along with their dissimilarities

Results


1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
Precision
0.00
0.05
0.10
0.15
0.20
Recall
HRERE-full
HRERE-naive
HRERE-base
Weston
PCNN+ATT
CNN+ATT

Cont.

P@N(%)	10%	30%	50%
Weston	79.3	68.6	60.9
HRERE-base	81.8	70.1	60.7
HRERE-naive	83.6	74.4	65.7
HRERE-full	**86.1**	**76.6**	**68.1**

Table: P@N of **Weston** and variants of our proposed framework.

Case Study

Much of the **middle east** tension stems from the sense that shiite power is growing, led by **Iran**.
relation: *contains* base: 0.311 naive: 0.864 **full**: **0.884**
Sometimes I rattle off the names of movie stars from **Omaha**: Fred Astaire, **Henry Fonda**, Nick Nolte . . .
relation: *place_of_birth* base: 0.109 naive: 0.605 **full**: **0.646**

Table: Some examples in NYT corpus and the predicted probabilities of the true relations.

Table of Contents

Conclusion

In this thesis, we

1. propose a single, much simpler and more elegant neural model that attempts FETC "end-to-end" without post-processing or ad-hoc features
2. propose a neural model with multi-level attention mechanisms for relation extraction
3. describe a neural framework for jointly learning heterogeneous representations from both text information and facts in an existing knowledge base to facilitate relation extraction

Q&A

Questions?