Towards Neural Information Extraction without Manual Annotated Data

Peng Xu

Department of Computing Science
University of Alberta
1. Introduction

2. Neural Fine-Grained Entity Type Classification (NFETC)

3. Neural Relation Extraction (NRE)

4. Incorporating Encoded Knowledge Information

5. Conclusion
Table of Contents

1 Introduction

2 Neural Fine-Grained Entity Type Classification (NFETC)

3 Neural Relation Extraction (NRE)

4 Incorporating Encoded Knowledge Information

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

2. Neural Fine-Grained Entity Type Classification (NFETC)

3. Neural Relation Extraction (NRE)

4. Incorporating Encoded Knowledge Information

5. Conclusion
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
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.)

<table>
<thead>
<tr>
<th></th>
<th>FIGER(GOLD)</th>
<th>OntoNotes</th>
</tr>
</thead>
<tbody>
<tr>
<td># types</td>
<td>113</td>
<td>89</td>
</tr>
<tr>
<td># raw testing mentions</td>
<td>563</td>
<td>8963</td>
</tr>
<tr>
<td>% testing mentions with single type-path</td>
<td>88.28</td>
<td>94.00</td>
</tr>
</tbody>
</table>
One kind of noise introduced by distant supervision is assigning labels that are *out-of-context*.
Another source of noise introduced by distant supervision is when the type is overly-specific for the context.
## Typical FETC Methods

<table>
<thead>
<tr>
<th></th>
<th>Attentive</th>
<th>AFET</th>
<th>LNR</th>
<th>A&amp;A</th>
</tr>
</thead>
<tbody>
<tr>
<td>without manual features</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>use attentive neural network</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>adopt single label setting</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>handle <em>out-of-context</em> noise</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>handle <em>overly-specific</em> noise</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
</tbody>
</table>

2. AFET: Ren *et al.* (2016a)
3. LNR: Ren *et al.* (2016b)
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.

<table>
<thead>
<tr>
<th></th>
<th>Attentive</th>
<th>AFET</th>
<th>LNR</th>
<th>A&amp;A</th>
<th>NFETC</th>
</tr>
</thead>
<tbody>
<tr>
<td>without manual features</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>use attentive neural network</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>adopt single label setting</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>handle <em>out-of-context</em> noise</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>handle <em>overly-specific</em> noise</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
</tbody>
</table>
Neural Architecture

At the Australian Open, she lost in the first round to #11 seed.
A Simple Yet Effective Variant of Cross-Entropy

Traditional cross-entropy loss:

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

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:

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

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.

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

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

\[ p^*(\hat{y}) = p(\hat{y}) + \beta \sum_{t \in \Gamma} p(t) \]  \hspace{1cm} (3)

where \( \Gamma \) is the set of ancestor types along the type-path of \( \hat{y} \), \( \beta \) 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

---

1Hierarchical 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

<table>
<thead>
<tr>
<th></th>
<th>FIGER(GOLD)</th>
<th>OntoNotes</th>
</tr>
</thead>
<tbody>
<tr>
<td># types</td>
<td>113</td>
<td>89</td>
</tr>
<tr>
<td># raw training mentions</td>
<td>2009898</td>
<td>253241</td>
</tr>
<tr>
<td># raw testing mentions</td>
<td>563</td>
<td>8963</td>
</tr>
<tr>
<td>% filtered training mentions</td>
<td>64.46</td>
<td>73.13</td>
</tr>
<tr>
<td>% filtered testing mentions</td>
<td>88.28</td>
<td>94.00</td>
</tr>
<tr>
<td>Max hierarchy depth</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>FIGER(GOLD)</th>
<th></th>
<th></th>
<th>OntoNotes</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Strict Acc.</td>
<td>Macro F1</td>
<td>Micro F1</td>
<td>Strict Acc.</td>
<td>Macro F1</td>
</tr>
<tr>
<td>Attentive</td>
<td>59.68</td>
<td>78.97</td>
<td>75.36</td>
<td></td>
<td>51.74</td>
<td>70.98</td>
</tr>
<tr>
<td>AFET</td>
<td>53.3</td>
<td>69.3</td>
<td>66.4</td>
<td></td>
<td>55.1</td>
<td>71.1</td>
</tr>
<tr>
<td>LNR+FIGER</td>
<td>59.9</td>
<td>76.3</td>
<td>74.9</td>
<td></td>
<td>57.2</td>
<td>71.5</td>
</tr>
<tr>
<td>A&amp;A</td>
<td>65.8</td>
<td>81.2</td>
<td>77.4</td>
<td></td>
<td>52.2</td>
<td>68.5</td>
</tr>
<tr>
<td><strong>NFETC(f)</strong></td>
<td>57.9 ± 1.3</td>
<td>78.4 ± 0.8</td>
<td>75.0 ± 0.7</td>
<td></td>
<td>54.4 ± 0.3</td>
<td>71.5 ± 0.4</td>
</tr>
<tr>
<td><strong>NFETC-hier(f)</strong></td>
<td>68.0 ± 0.8</td>
<td><strong>81.4 ± 0.8</strong></td>
<td>77.9 ± 0.7</td>
<td></td>
<td>59.6 ± 0.2</td>
<td><strong>76.1 ± 0.2</strong></td>
</tr>
<tr>
<td><strong>NFETC(r)</strong></td>
<td>56.2 ± 1.0</td>
<td>77.2 ± 0.9</td>
<td>74.3 ± 1.1</td>
<td></td>
<td>54.8 ± 0.4</td>
<td>71.8 ± 0.4</td>
</tr>
<tr>
<td><strong>NFETC-hier(r)</strong></td>
<td><strong>68.9 ± 0.6</strong></td>
<td><strong>81.9 ± 0.7</strong></td>
<td><strong>79.0 ± 0.7</strong></td>
<td></td>
<td><strong>60.2 ± 0.2</strong></td>
<td><strong>76.4 ± 0.1</strong></td>
</tr>
</tbody>
</table>

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(r)**: with variant of cross-entropy and hierarchical loss normalization trained on raw data
### Case Study

<table>
<thead>
<tr>
<th>Test Sentence</th>
<th>Ground Truth</th>
<th>Prediction (w/o HLN)</th>
<th>Prediction (w HLN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: <strong>Hopkins</strong> said four fellow elections is curious, considering the ...</td>
<td>Person</td>
<td>Politician</td>
<td>Person</td>
</tr>
<tr>
<td>S2: ...for WiFi communications across all the SD cards.</td>
<td>Product</td>
<td>Software</td>
<td>Product</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Sentence</th>
<th>Ground Truth</th>
<th>Prediction (original CE)</th>
<th>Prediction (improved CE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3: <strong>ASC</strong> Director Melvin Taing said that because the commission is ...</td>
<td>Organization</td>
<td>Title</td>
<td>Organization</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Sentence</th>
<th>Ground Truth</th>
<th>Failed Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4: A handful of professors in the <strong>UW</strong> Department of Chemistry ...</td>
<td>Educational Institution</td>
<td>Organization</td>
</tr>
<tr>
<td>S5: Work needs to be done and, in <strong>Washington state</strong>, ...</td>
<td>Province</td>
<td>City</td>
</tr>
</tbody>
</table>
Table of Contents

1  Introduction

2  Neural Fine-Grained Entity Type Classification (NFETC)

3  Neural Relation Extraction (NRE)

4  Incorporating Encoded Knowledge Information

5  Conclusion
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.

(Obama, alma mater, Harvard Law School)

In 1988 Obama enrolled in Harvard Law School, where he was the first black president of the Harvard Law Review.

(Obama, president of, Harvard Law Review)
- 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
Table of Contents

1 Introduction

2 Neural Fine-Grained Entity Type Classification (NFETC)

3 Neural Relation Extraction (NRE)

4 Incorporating Encoded Knowledge Information

5 Conclusion
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.
The language representation is learnt based on the input set of sentences $S$ for each entity pair with the neural architecture described in the previous section.

With the language representation, we can get the probability $p(r|S)$. 
Following the score function $\phi$ 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:

\[
J_L = - \frac{1}{N} \sum_{i=1}^{N} \log p(r_i | S_i)
\]

\[
J_G = - \frac{1}{N} \sum_{i=1}^{N} \log p(r_i | (h_i, t_i))
\]

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

\[
J_D = - \frac{1}{N} \sum_{i=1}^{N} p(r_i^* | S_i)
\]

where \( r_i^* = \arg \max p(r | (h_i, t_i)). \)
We form the joint optimization problem for model parameters as

$$\min_{\Theta} J = J_L + J_G + J_D + \lambda \|\Theta\|_2^2$$

where $\Theta$ 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 $\Theta$. 

Model Learning
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:
  - $\text{HRERE}$-base: basic neural model with local loss only
  - $\text{HRERE}$-naive: neural model with both local and global loss but without the dissimilarities
  - $\text{HRERE}$-full: neural model with both local and global loss along with their dissimilarities
Results
### Table: P@N of Weston and variants of our proposed framework.

<table>
<thead>
<tr>
<th>P@N(%)</th>
<th>10%</th>
<th>30%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weston</td>
<td>79.3</td>
<td>68.6</td>
<td>60.9</td>
</tr>
<tr>
<td>HRERE-base</td>
<td>81.8</td>
<td>70.1</td>
<td>60.7</td>
</tr>
<tr>
<td>HRERE-naive</td>
<td>83.6</td>
<td>74.4</td>
<td>65.7</td>
</tr>
<tr>
<td>HRERE-full</td>
<td><strong>86.1</strong></td>
<td><strong>76.6</strong></td>
<td><strong>68.1</strong></td>
</tr>
</tbody>
</table>
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

1. Introduction
2. Neural Fine-Grained Entity Type Classification (NFETC)
3. Neural Relation Extraction (NRE)
4. Incorporating Encoded Knowledge Information
5. Conclusion
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
Questions?