AFET: Automatic Fine-Grained Entity Typing by Hierarchical Partial-Label Embedding

Xiang Ren* (UIUC), Wenqi He* (UIUC), Meng Qu (UIUC), Lifu Huang (RPI), Heng Ji (RPI), Jiawei Han (UIUC)
Assigning type labels to mentions of entities in text:

- Type labels for a mention constitutes a “type-path” in a given type hierarchy (NOT necessarily ending in a leaf node).

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
<th>Type-path</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Republican presidential candidate <em>Donald Trump</em> spoke during a campaign event in Rock Hill.</td>
<td><strong>Person → politician</strong></td>
</tr>
<tr>
<td>S2</td>
<td><em>Donald Trump</em>’s company has threatened to withhold up to $1 billion of investment if the U.K. government decides to ban his entry into the country.</td>
<td><strong>Person → businessman</strong></td>
</tr>
<tr>
<td>S3</td>
<td>In <em>Trump</em>’s TV reality show, “The Apprentice”, 16 people competed for a job.</td>
<td><strong>Person → artist → actor</strong></td>
</tr>
</tbody>
</table>

Text corpus with mentions of entities (bold):

- *Republican presidential candidate Donald Trump*
- *Donald Trump*’s company
- *Trump*’s TV reality show, “The Apprentice”
Fine-grained Entity Typing: Applications

- Critical task for bringing structures to unstructured text data
  - Information network construction

- A primitive step in many NLP tasks
  - Coreference resolution, relation extraction, slot filing, ...

- Knowledge base completion

- Question answering

- etc.
Fine-Grained Typing with Distant Supervision

- Manually annotating training corpora with **100+ types**
- Expensive & Error-prone
- **Current practice**: using distant supervision to automatically label training corpora

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Republican presidential candidate <em>Donald Trump</em> spoke during a campaign event in Rock Hill.</td>
</tr>
</tbody>
</table>

Automatically typed mention in training corpus

1. **Mention**: “*Donald Trump*”; **Context**: S1; **Candidate Types**: {person, politician, businessman, artist, actor}
Problem Definition

- **Input:**
  - (1) Automatically labeled training corpus $D$
  - $\rightarrow$ (mention, context, candidate type labels) triples
  - (2) Target type hierarchy $T$
  - (3) Knowledge base (KB), and entity-type facts $\rightarrow$ (entity, type) tuples
  - (4) Test data $D_T$ $\rightarrow$ (mention, context) tuples
  - short text inputs: sentences, tweets, user queries, etc.

- **Output:**
  - Estimate *a single type-path* in $T$ for each test mention in $D_T$ based on the *mention itself* and its *context*

- **Non-goals:** Entity mention detection; Entity linking; Type hierarchy creation
Existing Studies

- **Fine-Grained Entity Typing**
  - Context-independent methods → not our focus
  - Context-dependent methods:
    1. Multi-label multi-class classification [FIGER-AAA’12, etc.]
    2. Hierarchical classification [HYENA-COLING’12]
    3. Joint mention-label embedding [WSABIE-ACL’15]
    4. Label propagation [ClusType-KDD’15, etc.]

- **Limitations**
  1. Except for 2, none of them capture semantic similarity between entity types
  2. All of them assume that automatically generated training labels are all “correct”
Existing Studies (cont.)

- **Partial Label Learning**
  - Assumes that “only one candidate label is correct” (1 among K is correct)
  - **Strategy 1**: Assumes equal contribution of each candidate label and average the outputs from all candidate labels for prediction [CLPL-JMLR’11]
  - **Strategy 2**: Progressively estimate true labels (as latent variables) using a EM-kind framework [PL-SVM-KDD’08]

- **Limitations**
  - Both cast the task as multi-label learning problem --- ignore the semantic similarity between entity types
  - Difficulty in handling high-dimensional feature space --- often the case in entity typing (0.4M+ features)
Challenges to Automate Fine-Grained Entity Typing

Are we using the training labels (by distant supervision) in a correct and comprehensive way?
Challenge I: Noisy Type Labels

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Republican presidential candidate Donald Trump spoke during a campaign event in Rock Hill.</td>
</tr>
<tr>
<td>S2</td>
<td>Donald Trump’s company has threatened to withhold up to $1 billion of investment if the U.K. government decides to ban his entry into the country.</td>
</tr>
<tr>
<td>S3</td>
<td>In Trump’s TV reality show, “The Apprentice”, 16 people competed for a job.</td>
</tr>
</tbody>
</table>

Donald Trump is mentioned in sentences S1-S3.

- Distant supervision
- Assign same types (blue region) to all the mentions
- Does not consider local contexts when assigning type labels
- Introduce label noise to the mentions

The types assigned to entity Trump include person, artist, actor, politician, businessman, while only {person, politician} are correct types for the mention “Trump” in S1
Challenge I: Noisy Type Labels (cont.)

- Current typing systems either **ignore this issue**
  - assume all the given type labels are “correct” labels

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Wiki</th>
<th>OntoNotes</th>
<th>BBN</th>
<th>NYT</th>
</tr>
</thead>
<tbody>
<tr>
<td># of target types</td>
<td>113</td>
<td>89</td>
<td>47</td>
<td>446</td>
</tr>
<tr>
<td>(1) noisy mentions (%)</td>
<td>27.99</td>
<td>25.94</td>
<td>22.32</td>
<td>51.81</td>
</tr>
<tr>
<td>(2a) sibling pruning (%)</td>
<td>23.92</td>
<td>16.09</td>
<td>22.32</td>
<td>39.26</td>
</tr>
<tr>
<td>(2b) min. pruning (%)</td>
<td>28.22</td>
<td>8.09</td>
<td>3.27</td>
<td>32.75</td>
</tr>
<tr>
<td>(2c) all pruning (%)</td>
<td>45.99</td>
<td>23.45</td>
<td>25.33</td>
<td>61.12</td>
</tr>
</tbody>
</table>

- Or use **simple pruning heuristics** to **delete** mentions with conflicting types
  - aggressive deletion of mentions $\rightarrow$ significant loss of training data

**The larger the target type set, the more severe the loss!**
Challenge II: Type Correlation

- Type correlation can be derived from the given type hierarchy or entity-type facts in KB
  - e.g., $\text{sim}(\text{singer}, \text{actor}) > \text{sim}(\text{politician}, \text{actor})$

- However, such information are ignored in existing methods → “every type label is treated equally”

- if **actor** is a true label for “**Trump**”, both **singer** and **politician** (as the false labels) will receive equal penalty

- In fact, penalty on **politician** should be larger than that on **singer**

- mentions of **actor** entities should be more similar to **singer** than to **politician** (instead of equally similar)
The Proposed Solution

- An Automatic Fine-Grained Entity Typing (AFET) Framework:
  - Jointly embed entity mentions (represented by text feature vectors) and type labels (in a given type hierarchy) into a low-dimensional space.
  - In that space, objects with similar types should be close to each other.
  - Essentially, we are learning embedding for each text feature and each type, and can use them to infer types for unseen mentions in test data.
  - Noisy labels are automatically handled in the framework.
The AFET Framework: Basic Idea I

- To handle noisy type labels

Model “clean” & “noisy” mentions using different objectives

- Clean mentions \( (m) \): \( \text{rank(true type} | m) < \text{rank(false type} | m) \)
- Noisy mentions \( (m') \): \( \text{rank(“best” candidate type} | m') < \text{rank(false type} | m') \)

- e.g., \( \text{rank(politician} | “Bernie Sanders”) < \text{rank(businessman} | “Bernie Sanders”) \)
- “Democratic presidential candidate \textit{Bernie Sanders} will make a campaign stop Friday at the University of Wisconsin”

- e.g., \( \text{rank( \{actor, artist, politician, businessman, ...\} | “Trump”) < rank(singer} | “Trump”) \)
- “In \textit{Trump}’s TV reality show, \textit{The Apprentice}, 16 people ...”
The AFET Framework: Basic Idea I

- To handle noisy type labels

Model “clean” & “noisy” mentions using different objectives

- **Clean mentions** \((m)\): \(\text{rank(\text{true type} | m)} < \text{rank(\text{false type} | m)}\)
- **Noisy mentions** \((m')\): \(\text{rank(“best” candidate type} | m') < \text{rank(\text{false type} | m')}\)

Progressively estimate the “best” candidate types for noisy mentions

- In each iteration, candidate types are scored based on \textit{currently estimated mention embeddings} & \textit{type embeddings} \(\rightarrow\) highest score as the “best”
- Text features used as side signal when learning mention embeddings
The AFET Framework: Basic Idea II

- To incorporate type correlation

  Adaptive margin between mention’s true type and false type, depending on the correlation between them

  - **Hierarchy-based correlation**: “close in the hierarchy” $\rightarrow$ more correlated
  - **KB-based correlation**: “sharing more entities in KB” $\rightarrow$ more correlated

- e.g., $\text{sim}(\text{singer}, \text{actor}) > \text{sim}(\text{politician}, \text{actor})$

- **Type distance**: $\gamma(X, Y) = 1 + 1 / [a + \text{sim}(X, Y)]$
  - $a$ is a given smoothing parameter
The AFET Framework: Basic Idea II

- To incorporate type correlation

Adaptive margin between mention’s true type and false type, depending on the correlation between them

• Require a large margin if the false type is less correlated to the true type (i.e., larger type distance between them)

- e.g., sim("Bernie Sanders", politician) - sim ("Bernie Sanders", businessman) > γ(politician, businessman) = 1
- e.g., sim("Bernie Sanders", politician) - sim ("Bernie Sanders", singer) > γ(politician, singer) = 10
- “Democratic presidential candidate Bernie Sanders will make a campaign stop ...”
The AFET Framework: Basic Idea II

To incorporate type correlation

Adaptive margin between mention’s true type and false type, depending on their correlation (distance)

• Require a large margin if the false type is less correlated to the true type (i.e., larger type distance between them)

For a clean mention, its fine-grained true type should be ranked higher than its coarse-grained true types

• e.g., rank(politician | “Bernie Sanders”) < rank(person | “Bernie Sanders”)
The AFET Framework: Overview

1. Use distant supervision to automatically label training corpus $D$

2. Extract text features for each mention in $D$; partition training mentions $M$ into clean set $M_c$ and noisy set $M_n$, based on the given type hierarchy $T$

3. Perform Hierarchical Partial-Label Embedding to learn mention embeddings and type embeddings jointly, by solving a proposed optimization problem

4. Predict type-path for each mention in test data $D_T$ following a top-down inference procedure
Text Features for Entity Mentions

- Features are extracted from:
  - (1) mention’s name string: *e.g.*, head token, POS tags, Brown Cluster of head token
  - (2) mention’s context in the sentence: *e.g.*, n-grams, dependency roles

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head Token</td>
<td>Syntactic head token of the mention</td>
<td>“HEAD_Turing”</td>
</tr>
<tr>
<td>POS</td>
<td>Tokens in the mention</td>
<td>“Turing”, “Machine”</td>
</tr>
<tr>
<td>Character</td>
<td>Part-of-Speech tag of tokens in the mention</td>
<td>“NN”</td>
</tr>
<tr>
<td>Word Shape</td>
<td>All character trigrams in the head of the mention</td>
<td>“:tu”, “tur”, ..., “ng:”</td>
</tr>
<tr>
<td>Length</td>
<td>Word shape of the tokens in the mention</td>
<td>“Aa” for “Turing”</td>
</tr>
<tr>
<td>Context</td>
<td>Number of tokens in the mention</td>
<td>“2”</td>
</tr>
<tr>
<td>Brown Cluster</td>
<td>Brown cluster ID for the head token (learned using D)</td>
<td>“CXT_B:Maserati ,”, “CXT_A:and the”</td>
</tr>
<tr>
<td>Dependency</td>
<td>Stanford syntactic dependency [16] associated with the head token</td>
<td>“4_1100”, “8_1101111”, “12_111011111111”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“GOV:nn”, “GOV:turing”</td>
</tr>
</tbody>
</table>

- **“Turing Machine”** is used as an example mention from the sentence:
  - “The band’s former drummer Jerry Fuchs—who was also a member of Maserati, Turing Machine and The Juan MacLean—died after falling down an elevator shaft.”.
The Joint Embedding Model

- Mapping entity mentions to the joint space
  - Mapping matrix $U$ (d-by-M)
    - Mention’s feature vector (dim = M)
    - Joint low-dimensional space for mentions and types
    - Mention embedding (dim = d)
  - Mapping entity types to the joint space
    - Mapping matrix $V$ (d-by-K)
      - Type indicator vector (dim = K)
      - Type embedding (dim = d)
    - $\Phi_M(m_i) = Um_i; \Phi_Y(y_k) =Vy_k$.
The Joint Embedding Model

- Mapping entity mentions to the joint space
  - Mapping matrix $U$ (d-by-M)
    - Mention’s feature vector (dim = M)
    - Joint low-dimensional space for mentions and types
  - Mapping entity types to the joint space
    - Mapping matrix $V$ (d-by-K)
      - Type indicator vector (dim = K)
      - Type embedding (dim = d)

$\Phi_M(m_i) = Um_i; \Phi_Y(y_k) = V y_k.$
Modeling Type Correlation by Hierarchy-Induced Margin

- When enforcing a margin between a mention’s true type and false type

\[ \Theta_{i,k,\bar{k}} = \max \left\{ 0, \gamma_{k,\bar{k}} - f_k(m_i) + f_{\bar{k}}(m_i) \right\} \]

Type distance between true type \( y_k \) and false type \( y_{\bar{k}} \)

For a mention \( (m_i) \) the margin between each of its true type labels \( (y_k) \) and each of its false type labels \( (y_{\bar{k}}) \) is required to be at least the type distance between them \( (\gamma_{k,\bar{k}}) \)

Similarity score between mention embedding & type embedding:

\[ f_k(m_i) = m_i^T U^T V y_k \]
Deriving Type Correlation and Distance

- **Hierarchy-based distance:** “close in the hierarchy” → more correlated
  
  \[ \gamma_{k,k'} = \text{length of the shortest path between } y_k \text{ and } y_{k'} \text{ in the given hierarchy (no path through root)} \]

- **KB-based correlation:** “sharing more entities in KB” → more correlated
  
  \[ w_{k,k'} = \left( \frac{\left| E_k \cap E_{k'} \right|}{|E_k|} + \frac{\left| E_k \cap E_{k'} \right|}{|E_{k'}|} \right) / 2 \]

- **KB-based distance:**
  
  \[ \gamma_{k,k'} = 1 + \frac{1}{[a + w_{k,k'}]} \]

Entity-Type Facts in KB:

- (Ben Affleck, actor)
- (Ben Affleck, director)
- (Woody Allen, actor)
- (Woody Allen, director)
- (J. K. Rowling, author)
- (Kobe Bryant, athlete)

Target Type Hierarchy (Tree):

- NO PATH \( \rightarrow \) \( \text{corr(person, location)} = 0 \)
- \( \text{root} \)
- product
- person
- location
- organization
- coach
- artist
- athlete
- author
- actor
- director

Diagram notes:

- \( \text{corr(actor, person)} = 1/(1+2) = 1/3 \)
- \( \text{corr(actor, author)} = 1/(1+2) = 1/3 \)
- \( \text{corr(actor, director)} = 1/(1+2) = 1/3 \)
Modeling Clean Mentions

- Extension of Weighted Approximate-Rank Pairwise (WARP) Loss

Margin-infused rank as the weight: if the true type is ranked lower among the false types, the weight becomes larger

\[
\text{rank}_{y_k}(f(m_i)) = \sum_{y_k \in \overline{Y_i}} \mathbb{1}(\gamma_{k,\bar{k}} + f_{k}(m_i) > f_{\bar{k}}(m_i))
\]

The loss for each clean mention:

\[
\ell_c(m_i, Y_i, \overline{Y_i}) = \sum_{y_k \in Y_i} \sum_{y_{\bar{k}} \in \overline{Y_i}} L\left[\text{rank}_{y_k}(f(m_i))\right] \Theta_{i,k,\bar{k}}
\]

For a clean mention, its “true types” should be ranked higher than all its “false types”
Modeling Clean Mentions

- Extension of Weighted Approximate-Rank Pairwise (WARP) Loss

  Margin-infused rank as the weight: if the true type is ranked lower among the false types, the weight becomes larger

  The loss for each clean mention:

  \[ \ell_c(m_i, \mathcal{Y}_i, \overline{\mathcal{Y}}_i) = \sum_{y_k \in \mathcal{Y}_i} \sum_{\overline{y}_k \in \overline{\mathcal{Y}}_i} L \left[ \text{rank}_{y_k} \left( f(m_i) \right) \right] \Theta_{i,k,\overline{k}} \]

  For every (true type, false type) pair of mention \( m_i \)

  \[ \Theta_{i,k,\overline{k}} = \max \left\{ 0, \gamma_{k,\overline{k}} - f_k(m_i) + f_{\overline{k}}(m_i) \right\} \]

  For a clean mention, its “true types” should be ranked higher than all its “false types"
Modeling Noisy Mentions

The loss for each noisy mention:

\[ \ell_n(m_i, \mathcal{Y}_i, \overline{\mathcal{Y}}_i) = L \left[ \text{rank}_{y_k^*} \left( f(m_i) \right) \right] \cdot \Omega_i \]

\[ \Omega_i = \max \left\{ 0, \gamma_{k^*, \bar{k}^*} - f_{k^*}(m_i) + f_{\bar{k}^*}(m_i) \right\} \]

For a mention \((m_i)\) the margin between each of its “best” candidate type \((y_{k^*})\) and its best false type \((y_{\bar{k}^*})\) is required to be at least the type distance between them \((\gamma_{k^*, \bar{k}^*})\).

For a noisy mention, its “best candidate type” should be ranked higher than all its “false types” (non-candidate types).
Modeling Noisy Mentions

The loss for each noisy mention:

\[
\ell_n(m_i, Y_i, \bar{Y}_i) = L \left[ \text{rank}_{y_k^*} \left( f(m_i) \right) \right] \cdot \Omega_i
\]

where

\[
\text{rank}_{y_k^*} \left( f(m_i) \right) = \sum_{y_k \in \bar{Y}_i} \mathbb{1}(\gamma_{k^*, \bar{k}} + f_{\bar{k}}(m_i) > f_{k^*}(m_i))
\]

Margin-infused rank as the weight: if the best candidate type is ranked lower among the false types, the weight becomes larger.

For a noisy mention, its “best candidate type” should be ranked higher than all its “false types” (non-candidate types).
The Optimization Problem for AFET

\[ \min_{U, V} \mathcal{O} = \mathcal{O}_c + \mathcal{O}_n \]

\[ = \sum_{m_i \in \mathcal{M}_c} \ell_c(m_i, \mathcal{Y}_i, \overline{\mathcal{Y}}_i) + \sum_{m_i \in \mathcal{M}_n} \ell_n(m_i, \mathcal{Y}_i, \overline{\mathcal{Y}}_i) \]

- Accounts for clean mentions \( \mathcal{M}_c \)
- Accounts for noisy mentions \( \mathcal{M}_n \)

- Minimize the objective \( \rightarrow \)
  - For clean mentions: true types are ranked higher than false types
  - For noisy mentions: best candidate type is ranked higher than false types
Model Learning

- Block-wise Coordinate Descent
- Can only do SGD but need to control the normalization cost
- Easy to parallelize by partitioning the mention set
- Discussions:
  - Since our type hierarchy size is acceptable (~100), we don’t need to do negative sampling → better performance
  - We don’t need to approximate the rank

Algorithm 2: Model Learning of AFET

Input: Feature vectors \( \{m_i\}_{i=1}^N \), Type vectors \( \{y_k\}_{k=1}^K \), learning rate \( \alpha \), normalization constant \( C \)

Output: feature embeddings \( U \), type embeddings \( V \)

Initialize: \( U \) and \( V \) as random matrices; while \( \partial \) in Eq. (11) not converge do

for \( m_i \in M_c \) do
  Compute the margin-infused rank for \( y_k \in Y_i \)
  Compute \( \partial \ell_{c,i} / \partial U \) using Eq. (12)
  Compute \( \partial \ell_{c,i} / \partial V \) using Eq. (15)
end

for \( m_i \in M_n \) do
  Compute the margin-infused rank for \( y_k \) *
  Compute \( \partial \ell_{n,i} / \partial U \) using Eq. (13)
  Compute \( \partial \ell_{n,i} / \partial V \) using Eq. (16)
end

\( U \leftarrow U - \alpha \cdot \left( \sum_{m_i \in M_c} \frac{\partial \ell_{c,i}}{\partial U} + \sum_{m_i \in M_n} \frac{\partial \ell_{n,i}}{\partial U} \right) \)

\( V \leftarrow V - \alpha \cdot \left( \sum_{m_i \in M_c} \frac{\partial \ell_{c,i}}{\partial V} + \sum_{m_i \in M_n} \frac{\partial \ell_{n,i}}{\partial V} \right) \)

Normalize the norms of \( U \) and \( V \) to \( C \)

end
Type Inference

- Perform top-down search in the given type hierarchy to estimate the correct type-path

---

### Algorithm 2: Type Inference

**Input:** candidate type sub-tree \( \{Y_i\} \), mention embeddings \( \{u_i\} \), type embeddings \( \{v_k\} \), threshold \( \eta \)

**Output:** estimated type-path \( \{Y_i^*\} \) for \( m_i \in \mathcal{M} \)

1. for \( m_i \in \mathcal{M} \) do
2.   Initialize: \( Y_i^* = \emptyset \), \( r \) as the root of \( Y \)
3.   while \( C_i(r) \neq \emptyset \) do
4.     \( r \leftarrow \arg \max_{y_k \in C_i(r)} s(u_i, v_k) \)
5.     if \( s(u_i, v_r) > \eta \) then
6.       Update the type-path: \( Y_i^* \leftarrow Y_i^* \cup \{r\} \)
7.     else
8.       return \( Y_i^* \) as the estimated type-path for \( m_i \)
9.   end
10. end

---

![Type Inference Diagram](image)
Experiment Setting

- **Datasets:**
  1. **Wiki:** 1.5M sentences sampled from ~780k Wikipedia articles
  2. **OntoNotes:** 13,109 news
  3. **BBN:** 2,311 Wall Street Journal articles

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Wiki</th>
<th>OntoNotes</th>
<th>BBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Types</td>
<td>113</td>
<td>89</td>
<td>47</td>
</tr>
<tr>
<td>#Documents</td>
<td>780,549</td>
<td>13,109</td>
<td>2,311</td>
</tr>
<tr>
<td>#Sentences</td>
<td>1.51M</td>
<td>143,709</td>
<td>48,899</td>
</tr>
<tr>
<td>#Training mentions</td>
<td>2.69M</td>
<td>223,342</td>
<td>109,090</td>
</tr>
<tr>
<td>#Ground-truth mentions</td>
<td>563</td>
<td>9,604</td>
<td>121,001</td>
</tr>
<tr>
<td>#Features</td>
<td>644,860</td>
<td>215,642</td>
<td>125,637</td>
</tr>
<tr>
<td>#Edges in graph</td>
<td>87M</td>
<td>5.9M</td>
<td>2.9M</td>
</tr>
</tbody>
</table>
Compared Methods

- **(1) Sib**: removes siblings types; **(2) Min**: removes types that appear only once in the document; **(3) All**: first performs Sib pruning then Min pruning;
- These three pruning strategies are combined with all baselines in the training stage
- **(1) FIGER**: it adopts multi-label perceptron. **(2) HYENA**: it adopts hierarchical SVM.
- **(3) DeepWalk**: embedding a homogeneous graph with binary edges; **(4) LINE**: second-order LINE for feature-label bipartite graph; **(5) WSABIE**: adopts WARP loss for feature-label bipartite graph; **(6) PTE**: applied PTE joint training algorithm
- **(7) PL-SVM**: uses a margin-based loss to handle label noise. **(8) CLPL**: uses a linear model to encourage large average scores for candidate types.
- For AFET, we compare **(1)AFET**: adopts KB-based type correlation; **(2)AFET-CoH**: adopts hierarchy-based type correlation; **(3)AFET-NoPartial**: does not model noisy mentions using the partial-label loss.
- To do: **(1) DeepNeuralNetwork** [IJCAI’15]. **(2) ClusType**.
Performance Comparison on Fine-Grained Typing

<table>
<thead>
<tr>
<th>Typing Method</th>
<th>Wiki</th>
<th></th>
<th></th>
<th>OntoNotes</th>
<th></th>
<th></th>
<th>BBN</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>Ma-F1</td>
<td>Mi-F1</td>
<td>Acc</td>
<td>Ma-F1</td>
<td>Mi-F1</td>
<td>Acc</td>
<td>Ma-F1</td>
<td>Mi-F1</td>
</tr>
<tr>
<td>CLPL (Cour et al., 2011)</td>
<td>0.162</td>
<td>0.431</td>
<td>0.411</td>
<td>0.201</td>
<td>0.347</td>
<td>0.358</td>
<td>0.438</td>
<td>0.603</td>
<td>0.536</td>
</tr>
<tr>
<td>PL-SVM (Nguyen and Caruana, 2008)</td>
<td>0.428</td>
<td>0.613</td>
<td>0.571</td>
<td>0.225</td>
<td>0.455</td>
<td>0.437</td>
<td>0.465</td>
<td>0.648</td>
<td>0.582</td>
</tr>
<tr>
<td>FIGER (Ling and Weld, 2012)</td>
<td>0.474</td>
<td>0.692</td>
<td>0.655</td>
<td>0.369</td>
<td>0.578</td>
<td>0.516</td>
<td>0.467</td>
<td>0.672</td>
<td>0.612</td>
</tr>
<tr>
<td>FIGER-Min (Gillick et al., 2014)</td>
<td>0.453</td>
<td>0.691</td>
<td>0.631</td>
<td>0.373</td>
<td>0.570</td>
<td>0.509</td>
<td>0.444</td>
<td>0.671</td>
<td>0.613</td>
</tr>
<tr>
<td>HYENA (Yosef et al., 2012)</td>
<td>0.288</td>
<td>0.528</td>
<td>0.506</td>
<td>0.249</td>
<td>0.497</td>
<td>0.446</td>
<td>0.523</td>
<td>0.576</td>
<td>0.587</td>
</tr>
<tr>
<td>HYENA-Min</td>
<td>0.325</td>
<td>0.566</td>
<td>0.536</td>
<td>0.295</td>
<td>0.523</td>
<td>0.470</td>
<td>0.524</td>
<td>0.582</td>
<td>0.595</td>
</tr>
<tr>
<td>DeepWalk (Perozzi et al., 2014)</td>
<td>0.414</td>
<td>0.563</td>
<td>0.511</td>
<td>0.479</td>
<td>0.669</td>
<td>0.611</td>
<td>0.586</td>
<td>0.638</td>
<td>0.628</td>
</tr>
<tr>
<td>WSABIE (Yogatama et al., 2015)</td>
<td>0.373</td>
<td>0.565</td>
<td>0.521</td>
<td>0.404</td>
<td>0.580</td>
<td>0.527</td>
<td>0.623</td>
<td>0.675</td>
<td>0.687</td>
</tr>
<tr>
<td>LINE (Tang et al., 2015b)</td>
<td>0.181</td>
<td>0.480</td>
<td>0.499</td>
<td>0.436</td>
<td>0.634</td>
<td>0.578</td>
<td>0.576</td>
<td>0.687</td>
<td>0.690</td>
</tr>
<tr>
<td>PTE (Tang et al., 2015a)</td>
<td>0.405</td>
<td>0.575</td>
<td>0.526</td>
<td>0.436</td>
<td>0.630</td>
<td>0.572</td>
<td>0.604</td>
<td>0.684</td>
<td>0.695</td>
</tr>
<tr>
<td>AFET-noPartial</td>
<td>0.476</td>
<td>0.627</td>
<td>0.597</td>
<td>0.452</td>
<td>0.592</td>
<td>0.531</td>
<td>0.657</td>
<td>0.702</td>
<td>0.709</td>
</tr>
<tr>
<td>AFET-CorrH</td>
<td>0.433</td>
<td>0.583</td>
<td>0.551</td>
<td>0.521</td>
<td>0.680</td>
<td>0.609</td>
<td>0.636</td>
<td>0.706</td>
<td>0.715</td>
</tr>
<tr>
<td>AFET-EditDist</td>
<td>0.520</td>
<td>0.714</td>
<td>0.682</td>
<td>0.547</td>
<td>0.705</td>
<td>0.645</td>
<td>0.664</td>
<td>0.722</td>
<td>0.726</td>
</tr>
<tr>
<td>AFET</td>
<td>0.503</td>
<td>0.689</td>
<td>0.655</td>
<td>0.551</td>
<td>0.711</td>
<td>0.647</td>
<td>0.665</td>
<td>0.726</td>
<td>0.729</td>
</tr>
</tbody>
</table>

- AFET vs. AFET-noPartial → gain by modeling noisy mentions
- AFET-noPartial vs. WSABIE → gain by incorporating type correlation
### Case Study I: Type Prediction for Test Mentions

<table>
<thead>
<tr>
<th>Text</th>
<th>Ground Truth</th>
<th>FIGER</th>
<th>WSABIE</th>
<th>PTE</th>
<th>AFET</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Valley Federal Savings &amp; Loan Association</strong> said Imperial Corp. of America withdrew from regulators its application to buy five Valley Federal branches, leaving the transaction in limbo.</td>
<td>organization, company</td>
<td>organization</td>
<td>organization, company, broadcast</td>
<td>organization</td>
<td>organization, company</td>
</tr>
<tr>
<td>It's terrific for advertisers to know the reader]will be paying more, &quot; said Michael Drexler, <em>national media director</em> at Bozell Inc. ad agency</td>
<td>person, title</td>
<td>organization</td>
<td>organization, company, news</td>
<td>person</td>
<td>person, title</td>
</tr>
</tbody>
</table>
Case Study II: Robustness to Noisy Training Labels

<table>
<thead>
<tr>
<th>Text</th>
<th>Harry fondly remembers the “old” days of the early ‘70s, when people like his friend Travis would take a psychiatrist on a date to analyze what Travis was doing wrong.</th>
<th>NASA says it may decide by tomorrow whether another space walk will be needed to fix torn thermal blanket that blew a cockpit.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate types</td>
<td>organization, music, person, artist</td>
<td>person, artist, location, structure, organization, company, news company</td>
</tr>
<tr>
<td>WSABIE</td>
<td>organization</td>
<td>person, artist</td>
</tr>
<tr>
<td>PTE</td>
<td>person, artist, music</td>
<td>organization, company, news company</td>
</tr>
<tr>
<td>AFET</td>
<td>person</td>
<td>organization</td>
</tr>
</tbody>
</table>