Improved Relation Extraction with Feature-Rich Compositional Embedding Models

Mo Yu* Matt Gormley*
Mark Dredze

September 21, 2015
EMNLP

*Co-first authors
FCM or: How I Learned to Stop Worrying (about Deep Learning) and Love Features

Mo Yu* Matt Gormley*
Mark Dredze

September 21, 2015
EMNLP

*Co-first authors
Handcrafted Features

\[ p(y|x) \propto \exp(\Theta_y \cdot f) \]
Where do features come from?

- **Hand-crafted features**:
  - First word before M1
  - Second word before M1
  - Bag-of-words in M1
  - Head word of M1
  - Other word in between
  - First word after M2
  - Second word after M2
  - Bag-of-words in M2
  - Head word of M2
  - Bigrams in between
  - Words on dependency path
  - Country name list
  - Personal relative triggers
  - Personal title list
  - WordNet Tags
  - Heads of chunks in between
  - Path of phrase labels
  - Combination of entity types

- **Feature Engineering**:
  - Sun et al., 2011

- **Feature Learning**:
  - Zhou et al., 2005
Where do features come from?

**Feature Engineering**

- **hand-crafted features**
  - Sun et al., 2011
  - Zhou et al., 2005

**Feature Learning**

- **word embeddings**
  - Mikolov et al., 2013

Unsupervised learning

**CBOW model in Mikolov et al. (2013)**

```
cat: 0.11 0.23 ... -0.45

dog: 0.13 0.26 ... -0.52
```

- Look-up table
  - input (context words)
  - embedding
  - missing word

- Similar words, similar embeddings

- Input (context words)

- Classifier

- CBOW model in Mikolov et al. (2013)
Where do features come from?

- Feature Engineering
  - Sun et al., 2011
- Feature Learning
  - Zhou et al., 2005
  - Word embeddings: Mikolov et al., 2013
  - String embeddings: Socher, 2011
  - Convolutional Neural Networks (Collobert and Weston 2008)
  - The movie showed wars
  - Recursive Auto Encoder (Socher 2011)
  - The movie showed wars
Where do features come from?

- Feature Engineering
- Feature Learning

- Word embeddings: Mikolov et al., 2013
- Tree embeddings: Socher et al., 2013
- String embeddings: Socher, 2011
- String embeddings: Collobert & Weston, 2008

The [movie] showed [wars]
Where do features come from?

Hand-crafted features
- Sun et al., 2011
- Zhou et al., 2005

Word embedding features
- Turian et al., 2010
- Koo et al., 2008

Word embeddings
- Mikolov et al., 2013

Tree embeddings
- Socher et al., 2013
- Hermann & Blunsom, 2013

String embeddings
- Socher, 2011
- Collobert & Weston, 2008

Refine embedding features with semantic/syntactic info
Where do features come from?

Our model (FCM)

- Mikolov et al., 2013
- Collobert & Weston, 2008

Feature Learning

- Socher et al., 2013
- Hermann & Blunsom, 2013

Feature Engineering

- Sun et al., 2011
- Koo et al., 2008
- Turian et al., 2010
- Hermann et al., 2014

hand-crafted features

- Zhou et al., 2005

word embeddings

- Mikolov et al., 2013

word embedding features
Feature-rich Compositional Embedding Model (FCM)

Goals for our Model:
1. Incorporate semantic/syntactic structural information
2. Incorporate word meaning
3. Bridge the gap between feature engineering and feature learning – but remain as simple as possible
Feature-rich Compositional Embedding Model (FCM)

Per-word Features:

<table>
<thead>
<tr>
<th>Feature</th>
<th>( f_1 )</th>
<th>( f_2 )</th>
<th>( f_3 )</th>
<th>( f_4 )</th>
<th>( f_5 )</th>
<th>( f_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{on-path}(w_i)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>\text{is-between}(w_i)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>\text{head-of-M1}(w_i)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>\text{head-of-M2}(w_i)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>\text{before-M1}(w_i)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\text{before-M2}(w_i)</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

The \([\text{movie}]_{M1}\) I watched depicted \([\text{hope}]_{M2}\)
Feature-rich Compositional Embedding Model (FCM)

Per-word Features:

- on-path($w_i$)
- is-between($w_i$)
- head-of-$M_1$($w_i$)
- head-of-$M_2$($w_i$)
- before-$M_1$($w_i$)
- before-$M_2$($w_i$)

\[ f_5 \]

\[
\begin{array}{c}
1 \\
1 \\
0 \\
0 \\
0 \\
1 \\
\ldots \\
\end{array}
\]

\[
\begin{array}{c}
\text{- nil} \\
\text{- noun-other} \\
\text{- noun-person} \\
\text{- verb-percep.} \\
\text{- verb-comm.} \\
\text{- noun-other} \\
\end{array}
\]

The $\text{[movie]}_{M_1} \ I \ \text{watched} \ \text{depicted} \ \text{[hope]}_{M_2}$
Feature-rich Compositional Embedding Model (FCM)

Per-word Features: (with conjunction)

- on-path\( (w_i) \) & \( w_i = \text{"depicted"} \)
- is-between\( (w_i) \) & \( w_i = \text{"depicted"} \)
- head-of-M1\( (w_i) \) & \( w_i = \text{"depicted"} \)
- head-of-M2\( (w_i) \) & \( w_i = \text{"depicted"} \)
- before-M1\( (w_i) \) & \( w_i = \text{"depicted"} \)
- before-M2\( (w_i) \) & \( w_i = \text{"depicted"} \)

...
Feature-rich Compositional Embedding Model (FCM)

Per-word Features: (with soft conjunction)

- on-path($w_i$)
- is-between($w_i$)
- head-of-M1($w_i$)
- head-of-M2($w_i$)
- before-M1($w_i$)
- before-M2($w_i$)

...
Feature-rich Compositional Embedding Model (FCM)

Per-word Features: (with soft conjunction)

on-path\((w_i)\)
is-between\((w_i)\)
head-of-\(M_1\)(\(w_i)\)
head-of-\(M_2\)(\(w_i)\)
before-\(M_1\)(\(w_i)\)
before-\(M_2\)(\(w_i)\)

\[
\begin{array}{cccc}
1 & -.3 & .9 & .1 & -1 \\
1 & -.3 & .9 & .1 & -1 \\
0 & -.3 & .9 & .1 & -1 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
1 & -.3 & .9 & .1 & -1 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
-.3 & .9 & .1 & -1 \\
\end{array}
\]

\[e_{\text{depicted}}\]
Feature-rich Compositional Embedding Model (FCM)

\[ p(y|x) \propto \exp \left( \sum_{i=1}^{n} f_i \right) \]

And finally, exponentiates and renormalizes

Then takes the dot-product with a parameter tensor

Our full model sums over each word in the sentence

The [movie]_{M1} I watched depicted [hope]_{M2}
Features for FCM

• Let $M_1$ and $M_2$ denote the left and right entity mentions

• Our per-word Binary Features:
  ▪ head of $M_1$
  ▪ head of $M_2$
  ▪ in-between $M_1$ and $M_2$
  ▪ −2, −1, +1, or +2 of $M_1$
  ▪ −2, −1, +1, or +2 of $M_2$
  ▪ on dependency path between $M_1$ and $M_2$

• Optionally:
  Add the entity type of $M_1$, $M_2$, or both
FCM as a Neural Network

- Embeddings are (optionally) treated as model parameters
- A log-bilinear model
- We initialize, then fine-tune the embeddings

Binary features

Embeddings

Parameter tensor

\[ p(y|x) \]

\[ e_x \]

\[ f_1 \]

\[ h_1 \]

\[ e_{w_1} \]

\[ f_n \]

\[ h_n \]

\[ e_{w_n} \]
Baseline Model

\[ Y_{i,j} \]

- Multinomial logistic regression (*standard approach*)
- Bring in all the usual binary NLP features (Sun et al., 2011)
  - type of the left entity mention
  - dependency path between mentions
  - bag of words in right mention
  - …

\[ \exp(\Theta_y f) \]
Hybrid Model: Baseline + FCM

\[ p(y|x) \propto \exp(\Theta_y f(y|x)) \]

Product of Experts:

\[ p(y|x) = \frac{1}{Z(x)} p_{\text{Baseline}}(y|x) p_{\text{FCM}}(y|x) \]
Experimental Setup

**ACE 2005**

- **Data:** 6 domains
  - Newswire (nw)
  - Broadcast Conversation (bc)
  - Broadcast News (bn)
  - Telephone Speech (cts)
  - Usenet Newsgroups (un)
  - Weblogs (wl)

- **Train:** bn+nw (~3600 relations)
- **Dev:** ½ of bc
- **Test:** ½ of bc, cts, wl

- **Metric:** Micro F1
  (given entity mention)

---

**SemEval-2010 Task 8**

- **Data:** Web text

- **Train:**
- **Dev:**
- **Test:** Standard split from shared task

- **Metric:** Macro F1
  (given entity boundaries)
ACE 2005 Results

Baseline
FCM
Baseline+FCM

Broadcast Conversation
Conversational Telephone Speech Test Set
Weblogs
## SemEval-2010 Results

<table>
<thead>
<tr>
<th>Source</th>
<th>Classifier</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socher et al. (2012)</td>
<td>RNN</td>
<td>74.8</td>
</tr>
<tr>
<td>Socher et al. (2012)</td>
<td>MVRNN</td>
<td>79.1</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb</td>
<td>81.8</td>
</tr>
<tr>
<td>Rink and Harabagiu (2010)</td>
<td>SVM</td>
<td>82.2</td>
</tr>
<tr>
<td>Zeng et al. (2014)</td>
<td>CNN</td>
<td>82.7</td>
</tr>
<tr>
<td>Santos et al. (2015)</td>
<td>CR-CNN (log-loss)</td>
<td>82.7</td>
</tr>
<tr>
<td>Liu et al. (2015)</td>
<td>DepNN</td>
<td>82.8</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb (task-spec-emb)</td>
<td>82.8</td>
</tr>
</tbody>
</table>

*Best in SemEval-2010 Shared Task*
## SemEval-2010 Results

<table>
<thead>
<tr>
<th>Source</th>
<th>Classifier</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socher et al. (2012)</td>
<td>RNN</td>
<td>74.8</td>
</tr>
<tr>
<td>Socher et al. (2012)</td>
<td>MVRNN</td>
<td>79.1</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb</td>
<td>81.8</td>
</tr>
<tr>
<td>Rink and Harabagiu (2010)</td>
<td>SVM</td>
<td>82.2</td>
</tr>
<tr>
<td>Zeng et al. (2014)</td>
<td>CNN</td>
<td>82.7</td>
</tr>
<tr>
<td>Santos et al. (2015)</td>
<td>CR-CNN (log-loss)</td>
<td>82.7</td>
</tr>
<tr>
<td>Liu et al. (2015)</td>
<td>DepNN</td>
<td>82.8</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb (task-spec-emb)</td>
<td>82.8</td>
</tr>
<tr>
<td></td>
<td>FCM (log-linear)</td>
<td>81.4</td>
</tr>
<tr>
<td></td>
<td>FCM (log-bilinear)</td>
<td>83.0</td>
</tr>
</tbody>
</table>

**Best in SemEval-2010 Shared Task**
<table>
<thead>
<tr>
<th>Source</th>
<th>Classifier</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socher et al. (2012)</td>
<td>RNN</td>
<td>74.8</td>
</tr>
<tr>
<td>Socher et al. (2012)</td>
<td>MVRNN</td>
<td>79.1</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb</td>
<td>81.8</td>
</tr>
<tr>
<td>Rink and Harabagiu (2010)</td>
<td>SVM</td>
<td>82.2</td>
</tr>
<tr>
<td>Zeng et al. (2014)</td>
<td>CNN</td>
<td>82.7</td>
</tr>
<tr>
<td>Santos et al. (2015)</td>
<td>CR-CNN (log-loss)</td>
<td>82.7</td>
</tr>
<tr>
<td>Liu et al. (2015)</td>
<td>DepNN</td>
<td>82.8</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb (task-spec-emb)</td>
<td>82.8</td>
</tr>
<tr>
<td></td>
<td>FCM (log-linear)</td>
<td>81.4</td>
</tr>
<tr>
<td></td>
<td>FCM (log-bilinear)</td>
<td>83.0</td>
</tr>
</tbody>
</table>
## SemEval-2010 Results

<table>
<thead>
<tr>
<th>Source</th>
<th>Classifier</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socher et al. (2012)</td>
<td>RNN</td>
<td>74.8</td>
</tr>
<tr>
<td>Socher et al. (2012)</td>
<td>MVRNN</td>
<td>79.1</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb</td>
<td>81.8</td>
</tr>
<tr>
<td>Rink and Harabagiu (2010)</td>
<td>SVM</td>
<td>82.2</td>
</tr>
<tr>
<td>Xu et al. (2015)</td>
<td>SDP-LSTM</td>
<td>82.4</td>
</tr>
<tr>
<td>Zeng et al. (2014)</td>
<td>CNN</td>
<td>82.7</td>
</tr>
<tr>
<td>Santos et al. (2015)</td>
<td>CR-CNN (log-loss)</td>
<td>82.7</td>
</tr>
<tr>
<td>Liu et al. (2015)</td>
<td>DepNN</td>
<td>82.8</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb (task-spec-emb)</td>
<td>82.8</td>
</tr>
<tr>
<td>Xu et al. (2015)</td>
<td>SDP-LSTM (full)</td>
<td>83.7</td>
</tr>
<tr>
<td></td>
<td>FCM (log-linear)</td>
<td>81.4</td>
</tr>
<tr>
<td></td>
<td>FCM (log-bilinear)</td>
<td>83.0</td>
</tr>
</tbody>
</table>
### SemEval-2010 Results

<table>
<thead>
<tr>
<th>Source</th>
<th>Classifier</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socher et al. (2012)</td>
<td>RNN</td>
<td>74.8</td>
</tr>
<tr>
<td>Socher et al. (2012)</td>
<td>MVRNN</td>
<td>79.1</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb</td>
<td>81.8</td>
</tr>
<tr>
<td>Rink and Harabagiu (2010)</td>
<td>SVM</td>
<td>82.2</td>
</tr>
<tr>
<td>Xu et al. (2015)</td>
<td>SDP-LSTM</td>
<td>82.4</td>
</tr>
<tr>
<td>Zeng et al. (2014)</td>
<td>CNN</td>
<td>82.7</td>
</tr>
<tr>
<td>Santos et al. (2015)</td>
<td>CR-CNN (log-loss)</td>
<td>82.7</td>
</tr>
<tr>
<td>Liu et al. (2015)</td>
<td>DepNN</td>
<td>82.8</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb (task-spec-emb)</td>
<td>82.8</td>
</tr>
<tr>
<td>Xu et al. (2015)</td>
<td>SDP-LSTM (full)</td>
<td>83.7</td>
</tr>
<tr>
<td>FCM (log-linear)</td>
<td></td>
<td>81.4</td>
</tr>
<tr>
<td>FCM (log-bilinear)</td>
<td></td>
<td>83.0</td>
</tr>
<tr>
<td>FCM (log-bilinear)</td>
<td>(task-spec-emb)</td>
<td>83.7</td>
</tr>
</tbody>
</table>
## SemEval-2010 Results

<table>
<thead>
<tr>
<th>Source</th>
<th>Classifier</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socher et al. (2012)</td>
<td>RNN</td>
<td>74.8</td>
</tr>
<tr>
<td>Socher et al. (2012)</td>
<td>MVRNN</td>
<td>79.1</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb</td>
<td>81.8</td>
</tr>
<tr>
<td>Rink and Harabagiu (2010)</td>
<td>SVM</td>
<td>82.2</td>
</tr>
<tr>
<td>Xu et al. (2015)</td>
<td>SDP-LSTM</td>
<td>82.4</td>
</tr>
<tr>
<td>Zeng et al. (2014)</td>
<td>CNN</td>
<td>82.7</td>
</tr>
<tr>
<td>Santos et al. (2015)</td>
<td>CR-CNN (log-loss)</td>
<td>82.7</td>
</tr>
<tr>
<td>Liu et al. (2015)</td>
<td>DepNN</td>
<td>82.8</td>
</tr>
<tr>
<td>Hashimoto et al. (2015)</td>
<td>RelEmb (task-spec-emb)</td>
<td>82.8</td>
</tr>
<tr>
<td>Xu et al. (2015)</td>
<td>SDP-LSTM (full)</td>
<td>83.7</td>
</tr>
<tr>
<td>Santos et al. (2015)</td>
<td>CR-CNN (ranking-loss)</td>
<td>84.1</td>
</tr>
<tr>
<td></td>
<td>FCM (log-linear)</td>
<td>81.4</td>
</tr>
<tr>
<td></td>
<td>FCM (log-bilinear)</td>
<td>83.0</td>
</tr>
<tr>
<td></td>
<td>FCM (log-bilinear) (task-spec-emb)</td>
<td>83.7</td>
</tr>
</tbody>
</table>
Takeaways

FCM bridges the gap between feature engineering and feature learning

If you are allergic to deep learning:

– Try the FCM for your task: it is simple, easy-to- implement, and was shown to be effective for two relation benchmarks

If you are a deep learning expert:

– Inject the FCM (i.e. outer product of features and embeddings) into your fancy deep network
Questions?

Two open source implementations:

- **Java**: (Within the Pacaya framework)
  [GitHub](https://github.com/mgormley/pacaya)

- **C++**: (From our NAACL 2015 paper on LRFCM)
  [GitHub](https://github.com/Gorov/ERE_RE)