Model Combination for Event Extraction in BioNLP 2011

Sebastian Riedel,\textsuperscript{a} David McClosky,\textsuperscript{b} Mihai Surdeanu,\textsuperscript{b} Andrew McCallum,\textsuperscript{a} and Christopher D. Manning\textsuperscript{b}

\textsuperscript{a}University of Massachusetts at Amherst and \textsuperscript{b}Stanford University

BioNLP 2011 — June 24th, 2011
Previous work / Motivation

- BioNLP 2009: model combination led to 4% F1 improvement over best individual system (Kim et al., 2009)
Previous work / Motivation

- BioNLP 2009: model combination led to 4% F1 improvement over best individual system (Kim et al., 2009)
- Netflix challenge: winning entry relies on model combination (Bennett et al., 2007)
Previous work / Motivation

- BioNLP 2009: model combination led to 4% F1 improvement over best individual system (Kim et al., 2009)
- Netflix challenge: winning entry relies on model combination (Bennett et al., 2007)
- CoNLL 2007: winning entry relies on model combination (Hall et al., 2007)
Previous work / Motivation

- BioNLP 2009: model combination led to 4% F1 improvement over best individual system (Kim et al., 2009)
- Netflix challenge: winning entry relies on model combination (Bennett et al., 2007)
- CoNLL 2007: winning entry relies on model combination (Hall et al., 2007)
- CoNLL 2003: winning entry relies on model combination (Florian et al., 2003)

- Most of these use stacking — so do we

  Stacked model's output as features in stacking model
Previous work / Motivation

- BioNLP 2009: model combination led to 4% F1 improvement over best individual system (Kim et al., 2009)
- Netflix challenge: winning entry relies on model combination (Bennett et al., 2007)
- CoNLL 2007: winning entry relies on model combination (Hall et al., 2007)
- CoNLL 2003: winning entry relies on model combination (Florian et al., 2003)
- etc. etc. etc.
Previous work / Motivation

- BioNLP 2009: model combination led to 4% F1 improvement over best individual system (Kim et al., 2009)
- Netflix challenge: winning entry relies on model combination (Bennett et al., 2007)
- CoNLL 2007: winning entry relies on model combination (Hall et al., 2007)
- CoNLL 2003: winning entry relies on model combination (Florian et al., 2003)
- etc. etc. etc.
- Most of these use **stacking**—so do we
Previous work / Motivation

- BioNLP 2009: model combination led to 4% F1 improvement over best individual system (Kim et al., 2009)
- Netflix challenge: winning entry relies on model combination (Bennett et al., 2007)
- CoNLL 2007: winning entry relies on model combination (Hall et al., 2007)
- CoNLL 2003: winning entry relies on model combination (Florian et al., 2003)
- etc. etc. etc.
- Most of these use **stacking**—so do we
- **Stacked** model’s output as features in **stacking** model
Stacking Model

Maximize \( s(e, a, b) = \sum_i s_i^T(e_i) + \sum_{i,j} s_{i,j}^A(a_{i,j}) + \sum_{p,q} s_{p,q}^B(b_{p,q}) \)

under global constrains

\[
\begin{align*}
\text{Binding} & : s(\text{Binding}) = -0.1 \\
\text{Regulation} & : s(\text{Regulation}) = 3.2 \\
\text{Phosphor.} & : s(\text{Phosphor.}) = 0.5
\end{align*}
\]

phosphorylation of TRAF2 inhibits binding to the CD40 domain

\[
\begin{align*}
& s(\text{None}) = -2.2 \\
& s(\text{Theme}) = 0.2 \\
& s(\text{Cause}) = 1.3 \\
& b_{p,q} \\
& s(\cdots) = -2.2 \\
& s(\cdots) = 3.2
\end{align*}
\]
Scores

\[ s(\text{Regulation}) = 3.2 \]

\[
\begin{pmatrix}
1 \\
\vdots \\
1
\end{pmatrix}^T
\begin{pmatrix}
-2.1 \\
\vdots \\
1.3
\end{pmatrix}
\]

\( e = \text{Reg} \)

\( e = \text{Reg and } w = "\text{inhibit}" \)
Stacked Features

\[
s(\text{Regulation}) = 3.2
\]

\[
\left( \begin{array}{c}
1 \\
1 \\
\vdots \\
1 \\
\end{array} \right)^\top \left( \begin{array}{c}
-2.1 \\
1.2 \\
\vdots \\
1.3 \\
\end{array} \right)
\]

- \(e = \text{Reg}\)
- \(e = \text{Reg and } y = \text{Reg}\)
- \(e = \text{Reg and } w = \text{”inhibit”}\)
Stacked model

- Stanford Event Parsing system

- Recall: Four different decoders:
  (1st, 2nd-order features) × (projective, non-projective)

- Only used the parser for stacking (1-best outputs)

- Different segmentation/tokenization

- Different trigger detection
Performance of individual components

| System     | $F_1$ |
|------------|-------|
| UMass      | 54.8  |

(Genia development section, Task 1)
## Performance of individual components

| System            | $F_1$ |
|-------------------|-------|
| UMass             | 54.8  |
| Stanford (1N)     | 49.9  |
| Stanford (1P)     | 49.0  |
| Stanford (2N)     | 46.5  |
| Stanford (2P)     | 49.5  |

(Genia development section, Task 1)
Performance of individual components

| System            | $F_1$ | with reranker |
|-------------------|-------|---------------|
| UMass             | 54.8  | —             |
| Stanford (1N)     | 49.9  | 50.2          |
| Stanford (1P)     | 49.0  | 49.4          |
| Stanford (2N)     | 46.5  | 47.9          |
| Stanford (2P)     | 49.5  | 50.5          |

(Genia development section, Task 1)
Model combination strategies

| System                               | $F_1$  |
|--------------------------------------|--------|
| UMass                                | 54.8   |
| Stanford (2P, reranked)              | 50.5   |
| UMass (1P)                           | 55.7   |
| UMass (2P)                           | 55.7   |
| UMass (all)                          | 55.9   |

(Genia development section, Task 1)
## Model combination strategies

| System                           | $F_1$ |
|----------------------------------|-------|
| UMass                            | 54.8  |
| Stanford (2P, reranked)          | 50.5  |
| Stanford (all, reranked)         | 50.7  |

(Genia development section, Task 1)
## Model combination strategies

| System                      | $F_1$ |
|-----------------------------|-------|
| UMass                       | 54.8  |
| Stanford (2P, reranked)     | 50.5  |
| Stanford (all, reranked)    | 50.7  |
| UMass←2N                    | 54.9  |
| UMass←1N                    | 55.6  |
| UMass←1P                    | 55.7  |
| UMass←2P                    | 55.7  |

(Genia development section, Task 1)
Model combination strategies

| System                          | $F_1$ |
|--------------------------------|-------|
| UMass                          | 54.8  |
| Stanford (2P, reranked)        | 50.5  |
| Stanford (all, reranked)       | 50.7  |
| UMass←2N                       | 54.9  |
| UMass←1N                       | 55.6  |
| UMass←1P                       | 55.7  |
| UMass←2P                       | 55.7  |
| UMass←all                      | 55.9  |

(Genia development section, Task 1)
Model combination strategies

| System                                      | $F_1$ |
|---------------------------------------------|-------|
| UMass                                       | 54.8  |
| Stanford (2P, reranked)                     | 50.5  |
| Stanford (all, reranked)                    | 50.7  |
| UMass←2N                                    | 54.9  |
| UMass←1N                                    | 55.6  |
| UMass←1P                                    | 55.7  |
| UMass←2P                                    | 55.7  |
| UMass←all (FAUST)                           | 55.9  |

(Genia development section, Task 1)
## Ablation analysis for stacking

| System                    | $F_1$ |
|---------------------------|-------|
| UMass                     | 54.8  |
| Stanford (2P, reranked)   | 50.5  |
| UMass←all                 | 55.9  |

*(Genia development section, Task 1)*
Ablation analysis for stacking

| System                        | $F_1$ |
|-------------------------------|-------|
| UMass                         | 54.8  |
| Stanford (2P, reranked)       | 50.5  |
| UMass←all                    | **55.9** |
| UMass←all (triggers)          | 54.9  |
| UMass←all (arguments)        | 55.1  |

(Genia development section, Task 1)
Conclusions

- Stacking: easy, effective method of model combination
- Variability in models critical for success
- Tree structure best provided by projective decoder
- Incorporated in UMass model via 2P stacking
- Future work: Incorporate projectivity constraint directly

Questions?
Conclusions

- Stacking: easy, effective method of model combination
  - ...even if base models differ significantly in performance

Future work: Incorporate projectivity constraint directly
Conclusions

- Stacking: easy, effective method of model combination
  - ...even if base models differ significantly in performance

- Variability in models critical for success
Conclusions

- Stacking: easy, effective method of model combination
  - ...even if base models differ significantly in performance

- Variability in models critical for success

- Tree structure best provided by projective decoder
Conclusions

- Stacking: easy, effective method of model combination
  - ...even if base models differ significantly in performance

- Variability in models critical for success

- Tree structure best provided by projective decoder
  - Incorporated in UMass model via 2P stacking
Conclusions

- Stacking: easy, effective method of model combination
  - ...even if base models differ significantly in performance

- Variability in models critical for success

- Tree structure best provided by projective decoder
  - Incorporated in UMass model via 2P stacking

- Future work: Incorporate projectivity constraint directly

Questions?
Backup slides
Stacked Features

\[
\begin{pmatrix}
1 \\
1 \\
\vdots \\
1
\end{pmatrix}^T
\begin{pmatrix}
-2.1 \\
1.2 \\
\vdots \\
1.3
\end{pmatrix}
\]

\[
es(\text{Regulation}) = 3.2
\]

- \(e = \text{Reg}\)
- \(e = \text{Reg and } y = \text{Reg}\)
- \(e = \text{Reg and } w = "\text{inhibit}"\)

13
Conjoined Features

\[
s(\text{Regulation}) = 3.2
\]

\[
\begin{pmatrix}
1 \\
1 \\
\vdots \\
1 \\
1
\end{pmatrix}^T
\begin{pmatrix}
-2.1 \\
1.2 \\
\vdots \\
1.3 \\
3.2
\end{pmatrix}
\]

- \(e = \text{Reg}\)
- \(e = \text{Reg and } y = \text{Reg}\)
- \(e = \text{Reg and } w = "\text{inhibit}"\)
- \(e = \text{Reg and } w = "\text{inhibit}" \text{ and } y = \text{Reg}\)
## Results on Genia

| System          | Simple | Binding | Regulation | Total |
|-----------------|--------|---------|------------|-------|
| UMass           | 74.7   | 47.7    | 42.8       | 54.8  |
| Stanford 1N     | 71.4   | 38.6    | 32.8       | 47.8  |
| Stanford 1P     | 70.8   | 35.9    | 31.1       | 46.5  |
| Stanford 2N     | 69.1   | 35.0    | 27.8       | 44.3  |
| Stanford 2P     | 72.0   | 36.2    | 32.2       | 47.4  |
| UMass ← All     | 76.9   | 43.5    | 44.0       | 55.9  |
| UMass ← 1N      | 76.4   | 45.1    | 43.8       | 55.6  |
| UMass ← 1P      | 75.8   | 43.1    | 44.6       | 55.7  |
| UMass ← 2N      | 74.9   | 42.8    | 43.8       | 54.9  |
| UMass ← 2P      | 75.7   | 46.0    | 44.1       | 55.7  |
| UMass ← All (triggers) | 76.4 | 41.2    | 43.1       | 54.9  |
| UMass ← All (arguments) | 76.1 | 41.7    | 43.6       | 55.1  |
## Results on Infectious Diseases

| System             | Rec  | Prec | $F_1$ |
|--------------------|------|------|-------|
| UMass              | 46.2 | 51.1 | 48.5  |
| Stanford 1N        | 43.1 | 49.1 | 45.9  |
| Stanford 1P        | 40.8 | 46.7 | 43.5  |
| Stanford 2N        | 41.6 | 53.9 | 46.9  |
| Stanford 2P        | 42.8 | 48.1 | 45.3  |
| UMass $\leftarrow$ All | 47.6 | 54.3 | 50.7  |
| UMass $\leftarrow$ 1N | 45.8 | 51.6 | 48.5  |
| UMass $\leftarrow$ 1P | 47.6 | 52.8 | 50.0  |
| UMass $\leftarrow$ 2N | 45.4 | 52.4 | 48.6  |
| UMass $\leftarrow$ 2P | 49.1 | 52.6 | 50.7  |
| UMass $\leftarrow$ 2P (conjoined) | 48.0 | 53.2 | 50.4  |
## Results on test

|                | UMass | UMass ↔ All |
|----------------|-------|-------------|
|                | Rec   | Prec | $F_1$  | Rec   | Prec | $F_1$  |
| **GE (Task 1)**| 48.5  | 64.1 | 55.2   | 49.4  | 64.8 | 56.0   |
| **GE (Task 2)**| 43.9  | 60.9 | 51.0   | 46.7  | 63.8 | 53.9   |
| **EPI (Full task)** | 28.1  | 41.6 | 33.5   | 28.9  | 44.5 | 35.0   |
| **EPI (Core task)** | 57.0  | 73.3 | 64.2   | 59.9  | 80.3 | 68.6   |
| **ID (Full task)**  | 46.9  | 62.0 | 53.4   | 48.0  | 66.0 | 55.6   |
| **ID (Core task)**  | 49.5  | 62.1 | 55.1   | 50.6  | 66.1 | 57.3   |