Event Detection and Domain Adaptation with Convolutional Neural Networks

Written by Thien Huu Nguyen and Ralph Grishman for ACL2015
Presented by Xia Cui for NLP@UoL
Overview

• First work on Event Detection using Convolutional Neural Networks
• First Research on Domain Adaptation using CNNs
• Overcome two fundamental limitations of the traditional feature-based approach
  • Complicated feature engineering for rich feature sets
  • Error propagation from the preceding stages which generate these features
Event Detection (ED)

- Identifying instances of specified types of events in text
- Associated with each *event mention*, identify *event triggers* and classify them into specific types
  - Event mention: event itself
  - Event trigger: evokes the event

- Example:
  
  A police officer was *killed* in New Jersey today.
  
  - “killed” is the trigger for the event “Die”
Challenges and Problems

• The same event might appear in the form of various trigger expressions and an expression might represent different events in different contexts → event extraction and event argument discovery

• Two problems in previous works
  • Feature selection is manual, requires linguistic intuition and domain expertise implying additional studies for new application domains and limiting the capacity to quickly adapt to new domains
  • Supervised NLP toolkits and resources for feature extraction might involve error (due to the imperfect nature or the performance loss of the toolkits on new domains)
Solution: Convolutional Neural Network (CNN)

• Automatically learns features from sentences, and minimises the dependence on supervised toolkits and resources for features

• Demonstrates significantly outperform traditional feature-based methods
  • Due to the capacity to mitigate the error propagation from the pre-processing modules for features
  • Due to the use of word embeddings to induce a more general representation for trigger candidates
Token Representations

• Event Detection $\rightarrow$ Multiclass classification problem

• $x = [x_{-w}, x_{-w+1}, \ldots, x_0, \ldots, x_{w-1}, x_w]$
  • $x$: event trigger (token)
  • $2w + 1$: fixed window size
  • $m_t \times (2w + 1)$: matrix size
  • $m_t$: dimensionality of the concatenated vectors of the tokens
Embedding Tables

- **Word Embedding Table**
  - Pretrained, semantic and syntactic properties

- **Position Embedding Table**
  - Initialized randomly, embed the relative distance $i$ of token $x_i$ to current token $x_0$

- **Entity Type Embedding Table**
  - Initialized randomly, using known entity type associated with each token in the sentence, **BIO annotation scheme** to assign entity type labels to each token in the trigger candidate
Gradients by back-propagation

Training: via stochastic gradient descent with shuffled mini-batches and AdaDelta update rule (Kim, 2014)
Experiment Settings

- Multiple window size (in convolution layer): \{2, 3, 4, 5\}
  - 150 feature maps for each window size
- Window size for the trigger: 31
- Dimensionality of pretrained word embeddings from word2vec: 300
- Dimensionality of position embeddings: 50
- Dimensionality of entity type embeddings: 50
- Parameters taken from Kim (2014):
  - Dropout rate $\rho = 0.5$, mini-batch size = 50, $l_2$ norms = 3
Experiment Datasets

- **ACE2005 corpus**
  - 34-class (33 types + 1 “none”)
  - Test set: 40 newswire articles (672 sentences)
  - Development set: 30 other documents (836 sentences)
  - Training set: remaining 529 documents (14,849 sentences)

Table 1: Performance on the Development Set.
Performance Comparison

| Methods                                                                 | P   | R   | F   |
|------------------------------------------------------------------------|-----|-----|-----|
| Sentence-level in Hong et al. (2011)                                   | 67.6| 53.5| 59.7|
| MaxEnt with local features in Li et al. (2013b)                       | 74.5| 59.1| 65.9|
| Joint beam search with local features in Li et al. (2013b)            | 73.7| 59.3| 65.7|
| Joint beam search with local and global features in Li et al. (2013b)  | 73.7| 62.3| 67.5|
| Cross-entity in Hong et al. (2011) †                                   | 72.9| 64.3| 68.3|
| CNN1: CNN without any external features                                | 71.9| 63.8| 67.6|
| CNN2: CNN augmented with entity types                                  | 71.8| 66.4| 69.0|

Table 2: Performance with Gold-Standard Entity Mentions and Types. † beyond sentence level.

• External features are the features generated from the supervised NLP modules and manual resources such as parsers, name taggers, entity mention extractors, gazetteer etc.

| Methods                                                                 | F   |
|------------------------------------------------------------------------|-----|
| Sentence level in Ji and Grishman (2008)                               | 59.7|
| MaxEnt with local features in Li et al. (2013b)                        | 64.7|
| Joint beam search with local features in Li et al. (2013b)             | 63.7|
| Joint beam search with local and global features in Li et al. (2013b)  | 65.6|
| CNN1: CNN without any external features                                | 67.6|

Table 3: Performance with Predicted Entity Mentions and Types.
Domain Adaptation

- **Goal:** develop techniques taking training data in some *source domain* and learning models that can work well on *target domains*
- **Datasets:** ACE2005 corpus
  - 6 **different domains:** broadcast conversation (bc), broadcast news (bn), telephone conversation (cts), newswire (nw), usenet (un) and webblog (wl)
  - **Source domain:** news (bn ∪ nw)
  - **Target domains:** bc, cts, wl (3 different domains)
  - **Development set:** half of bc
  - **Test set:** remaining half of bc
In-domain & out-of-domain Performance

| System                                           | In-domain(bn+nw) | bc   | cts   | wl   |
|--------------------------------------------------|------------------|------|-------|------|
|                                                  | P    | R    | F    | P    | R    | F    | P    | R    | F    |
| MaxEnt                                           | 74.5 | 59.4 | 66.0 | 70.1 | 54.5 | 61.3 | 66.4 | 49.9 | 56.9 |
| Joint beam search in Li et al. (2013b)           |      |      |      |      |      |      |      |      |      |
| Joint+Local                                      | 73.5 | 62.7 | 67.7 | 70.3 | 57.2 | 63.1 | 64.9 | 50.8 | 57.0 |
| Joint+Local+Global                               | 72.9 | 63.2 | 67.7 | 68.8 | 57.5 | 62.6 | 64.5 | 52.3 | 57.7 |
| CNN1                                             | 70.9 | 64.0 | 67.3 | 71.0 | 61.9 | 66.1†| 64.0 | 55.0 | 59.1 |
| CNN2                                             | 69.2 | 67.0 | **68.0** | 70.2 | 65.2 | **67.6†** | 68.3 | 58.2 | **62.8†** |

Table 4: In-domain (first column) and Out-of-domain Performance (columns two to four). Cells marked with † designate CNN models that significantly outperform ($p < 0.05$) all the reported feature-based methods on the specified domain.

The performance of feature-based systems MaxEnt, Joint+Local and Joint+Local+Global are obtained from the actual systems in Li et al. (2013b). CNN1 and CNN2 via 5-fold cross validation.