# Named Entity Recognition in Bio-medical Documents with an Optimized Tag-set

Han-Cheol Cho and Jun'ichi Tsujii
Tsujii Lab., The University of Tokyo

## 1. What is Named Entity Recognition ?

- ***Named Entity Recognition (NER)*** **task is to recognize and classify target entities in a given sentence**

… does not affect **proinflammatory cytokine** (**tumor necrosis factor-alpha**, **interleukin-6**, and **interferon-gamma**) release from …

- **NER problem can be transformed into *a sequential labeling problem***

… does/**O** not/**O** affect/**O** **proinflammatory**/**B-Protein** **cytokine**/**I-Protein** (/**O** **tumor**/**B-Protein** **necrosis**/**I-Protein** **factor-alpha**/**I-Protein** ,/**O** **interleukin-6**/**B-Protein** ,/**O** and/**O** **interferon-gamma**/**B-Protein** )/**O** release/**O** from/**O** …

## 2. TAG-SETS

- **Sequential labeling for NER is to annotate every word in a given sentence with *a pre-defined tag-set***
  **→ IO, IOB1, IOB2, IOE1, IOE2, …**
- **IOB2 tag-set has been a predominant tag-set for not only general NER but also Bio-medical NER**
- ***(Lev Ratinov and Dan Roth, 2009)* claimed that BILOU tag-set is better than IOB2 tag-set with the experiment results below**

| Tag-set | CoNLL03 | | MUC7 | |
|---|---|---|---|---|
| | Dev. | Test | Dev. | Test |
| IOB2 | **93.61%** | 89.15% | 86.76% | 85.15% |
| BILOU | 93.28% | **90.57%** | **88.09%** | **85.62%** |

## 3. Contradictory Result

- **However, BILOU tag-set shows a slightly lower performance compared to IOB2 tag-set in the experiment with BioNLP/NLPBA 2004 shared task corpus**

| Tag-set | NLPBA04 Test Data | | |
|---|---|---|---|
| | Recall | Precision | F1-score |
| IOB2 | **69.33%** | 67.13% | **68.21%** |
| BILOU | 68.99% | **67.16%** | 68.06% |

*** Complete Match Performance**

## 4. Data Sparseness Problem

- **The number of training data for each tag**

| | B | U | I | L | O |
|---|---|---|---|---|---|
| IOB2 | 51301 | | 58287 | | 382963 |
| BILOU | 29655 | 21646 | 28632 | 29655 | 382963 |

## 5. Taking a Different View Point

- **To reduce the data sparseness problem, divide O-tag into multiple tags depending on its position**

… does/**O-I** not/**O-I** affect/**O-I** **proinflammatory/B-Protein cytokine/I-Protein** (/**O-B** **tumor/B-Protein necrosis/I-Protein factor-alpha/I-Protein** , /**O-B** **interleukin-6/B-Protein** , /**O-B** and/**O-I** **interferon-gamma/B-Protein** ) /**O-B** release/**O-I** from/**O-I** …

## 6. Experiment Results

| Tag-set | NLPBA04 Test Data | | |
|---|---|---|---|
| | Recall | Precision | F1-score |
| IOB2 | 69.33% | 67.13% | 68.21% |
| $IO_2B2$ | 69.27% | 67.65% | 68.45% |
| BILOU | 68.99% | 67.16% | 68.06% |
| $BILO_4U$ | **69.57%** | **68.22%** | **68.89%** |

- **O-tag sub-categorization *primarily improves precision* (*0.52%* in IOB2 tag-set, *1.06%* in BILOU tag-set)**

## 7. Conclusion and Future Work

- **Conclusion**
  1. **BILOU tag-set does not outperform IOB2 tag-set in Bio-medical NER**
  2. **Sub-categorization of O-tag in both IOB2 and BILOU tag-sets improves NER performance, mainly precision**
- **Future Work**
  **Applying a Latent Variable Model (Xu Sun et al, 2009) that automatically exploits sub-categorized information**