Detecting Protein Names Using Syntactic Dependency Relations

Tong Wang
Computational Linguistics Group

Department of Computer Science
University of Toronto

Introduction

- Mining medical domain literature
  - named entity recognition (NER)
  - protein-protein interaction detection (PPID)

NE Recognition

- Protein names, DNA names, ...
- Classification problem (lexical level)
  - dictionary-based
  - machine learning approaches
  - mostly lexical features

NER Features

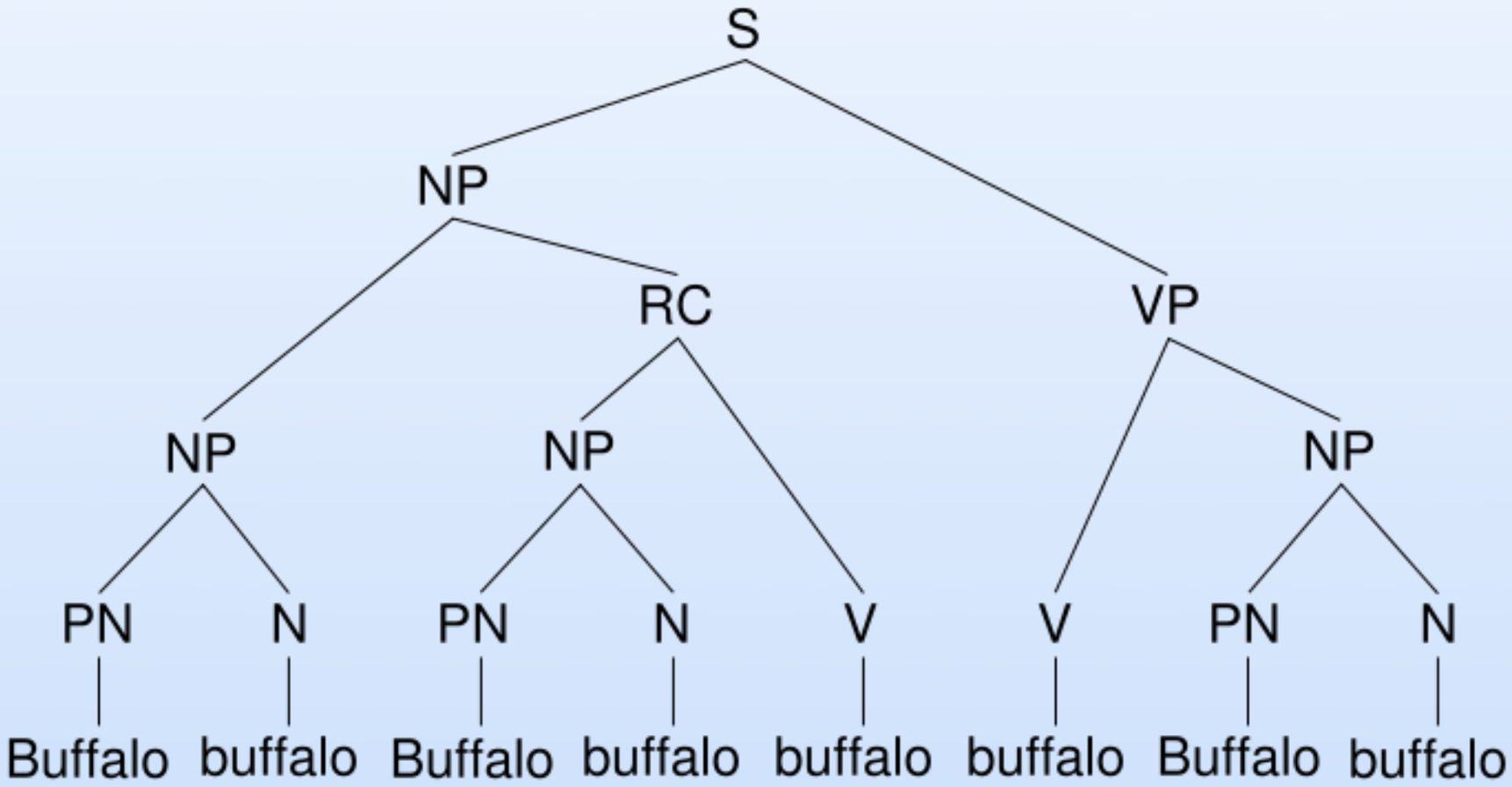
- BoW or n-gram
- POS
- orthographical
  - capitalization
  - digits/roman numbers
  - word shape: Kappa-2B => Aa_0A

PP Interaction Detection

- Is the sentence talking about PPI?
- Also classification problem (sentence level)
- Mainly machine learning approaches
- Lexical as well as *syntactical* features

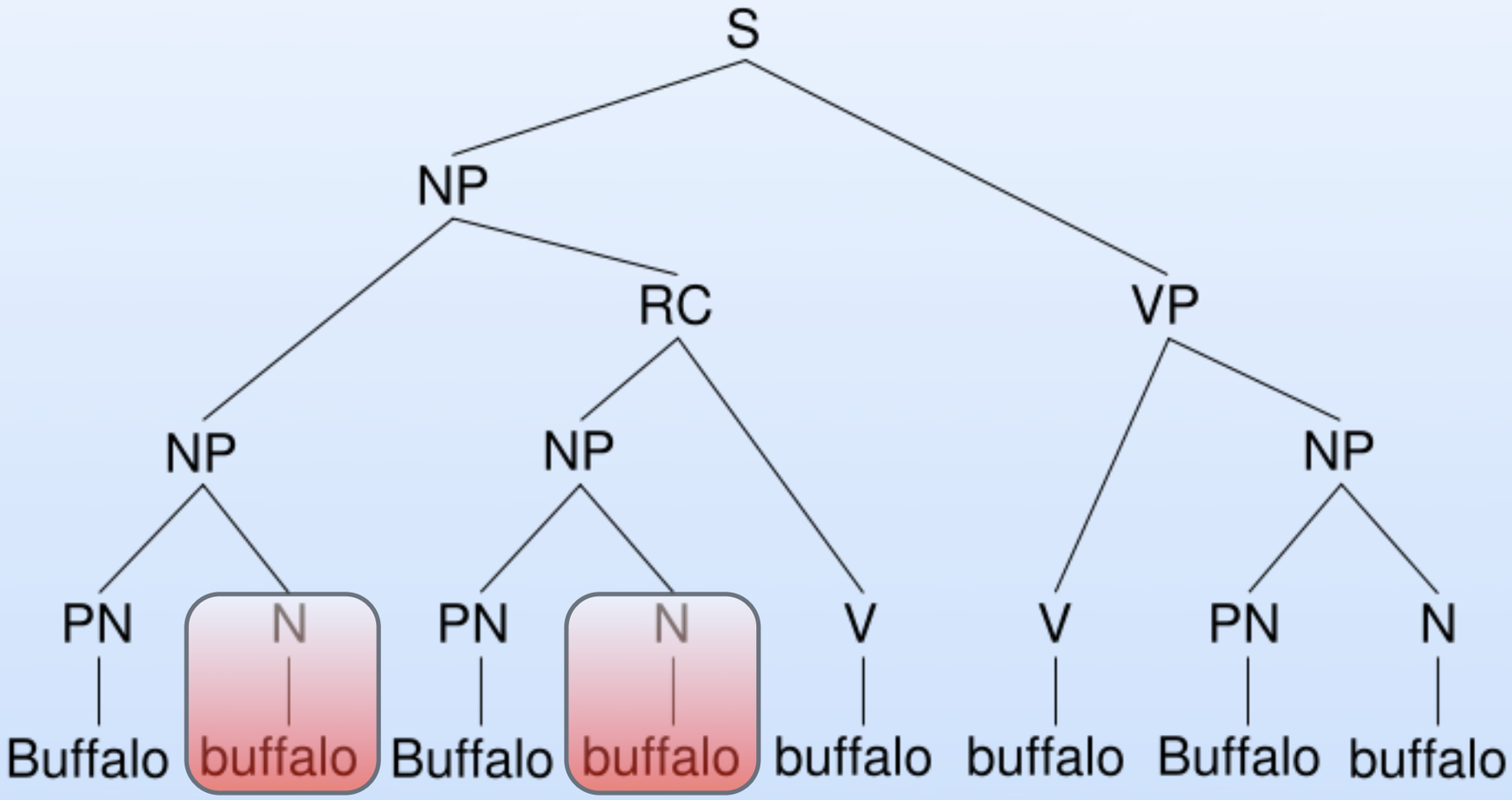
PPID Features

Parse Trees



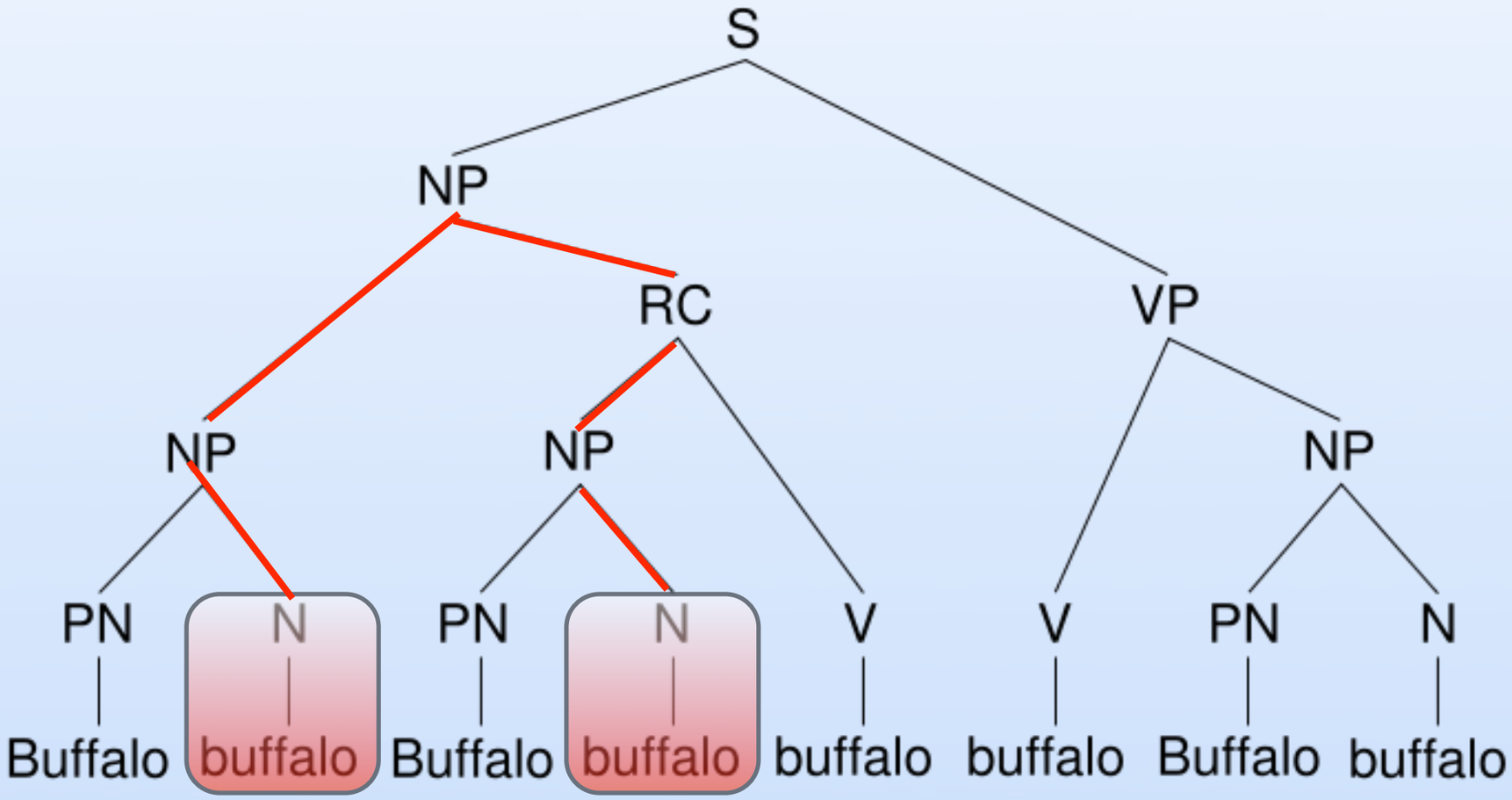
PPID Features

Parse Trees



PPID Features

Parse Trees



PPID Features

- Dependency relations (shallow parses)
  - HIV-2 caused ...
    - subj cause hiv-2
  - ... activation through CD28 ...
    - modpp activation cd28 through#4

Objective

- Features: combing what NER already has with dependency parsing
- Classifier: maximum entropy model

Data

- GENIA
  - 2,000 MEDLINE abstracts
  - 400,000 words, 100,000 annotations

```
<sentence id="m95369245-s1"><cons lex="IL-2_gene_expression" sem="G#other_name"><cons lex="IL-2_gene" sem="G#DNA_d
omain_or_region"><w c="NN" id="m95369245-w1">IL-2</w> <w c="NN" id="m95369245-w2">gene</w></cons> <w c="NN" id="m9
5369245-w3">expression</w></cons> <w c="CC" id="m95369245-w4">and</w> <cons lex="NF-kappa_B_activation" sem="G#oth
er_name"><cons lex="NF-kappa_B" sem="G#protein_molecule"><w c="NN" id="m95369245-w5">NF-kappa</w> <w c="NN" id="m9
5369245-w6">B</w></cons> <w c="NN" id="m95369245-w7">activation</w></cons> <w c="IN" id="m95369245-w8">through</w>
 <cons lex="CD28" sem="G#protein_molecule"><w c="NN" id="m95369245-w9">CD28</w></cons> <w c="VBZ" id="m95369245-w1
0">requires</w> <w c="JJ" id="m95369245-w11">reactive</w> <w c="NN" id="m95369245-w12">oxygen</w> <w c="NN" id="m9
5369245-w13">production</w> <w c="IN" id="m95369245-w14">by</w> <cons lex="5-lipoxygenase" sem="G#protein_molecule
"><w c="NN" id="m95369245-w15">5-lipoxygenase</w></cons><w c="." id="m95369245-w16">.</w></sentence>
```

IL-2 gene expression and F-Kappa B activation through CD28
requires reactive oxygen production by 5-lipoxygenase.

DNA-domain/region
Protein name

Department of Computer Science
University of Toronto

Data

- DepGENIA
  - GENIA with dependency parses
  - in a separate file, linked with sentence/ word ID

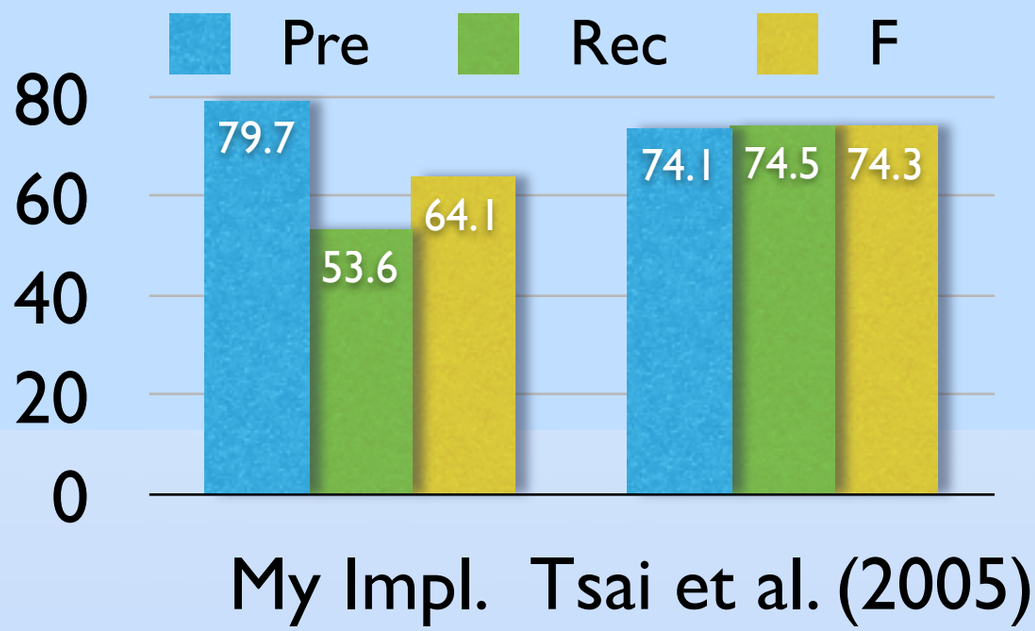
m95369245-s1	conj	m95369245-w7	m95369245-w4	activation	and	-
m95369245-s1	prep	m95369245-w9	m95369245-w8	cd28	through	-
m95369245-s1	modpp	m95369245-w7	m95369245-w9	activation	cd28	through#4
m95369245-s1	subj	m95369245-w10	m95369245-w7	require	activation	and#2
m95369245-s1	prep	m95369245-w15	m95369245-w14	5-lipoxygenase	by	-
m95369245-s1	modpp	m95369245-w13	m95369245-w15	production	5-lipoxygenase	by#8
m95369245-s1	obj	m95369245-w10	m95369245-w13	require	production	5-lipoxygenase#9
m95369245-s2	prep	m95369245-w22	m95369245-w18	receptor	of	-
m95369245-s2	modpp	m95369245-w17	m95369245-w22	activation	receptor	of#2

Workflow

- Data partitioning: 2–1–1
  - LOTS of DTD issues
- Implementing Tsai et al. (2005)
- Evaluation I
- *Adding dependency features*
- Evaluation II

Implementing Tsai et al. (2005)

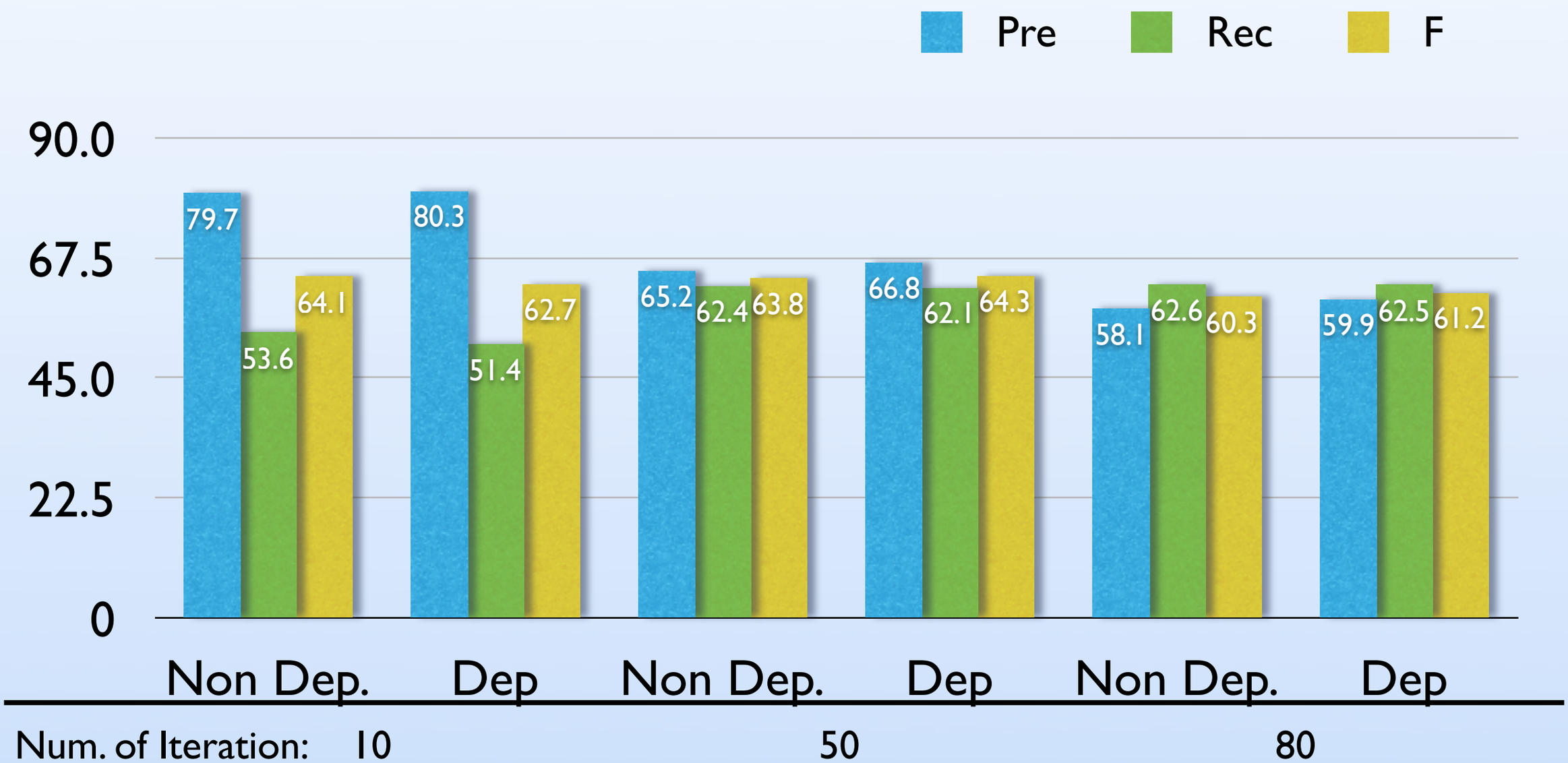
- Orthographical features: regex confusion
- Context feature: BoW or n-gram?
- POS tagging: MBT tagger vs. GENIA tagged
- Prefix/suffix length



Dependency Features Added (so far)

- modpp => head
- modpp => pp with sense number
- modpp => without sense number
- subj => verb
- obj => verb
- has_appositive, appositive

(Preliminary) Results



Department of Computer Science
University of Toronto

What's Next?

- More meaningful dependence features
- Clearing confusions on implementation of Tsai et al. (by comparing all alternatives)
- Error analysis on both parts
- Precision-recall curves

Thank you.

謝謝。

Department of Computer Science
University of Toronto