Challenges in Pronoun Resolution System for Biomedical Text

Ngan Nguyen, Jin-Dong Kim and Jun'ichi Tsujii

{ nltngan, jdkim, tsujii } @ is.s.u-tokyo.ac.jp

Department of Computer Science,

The University of Tokyo

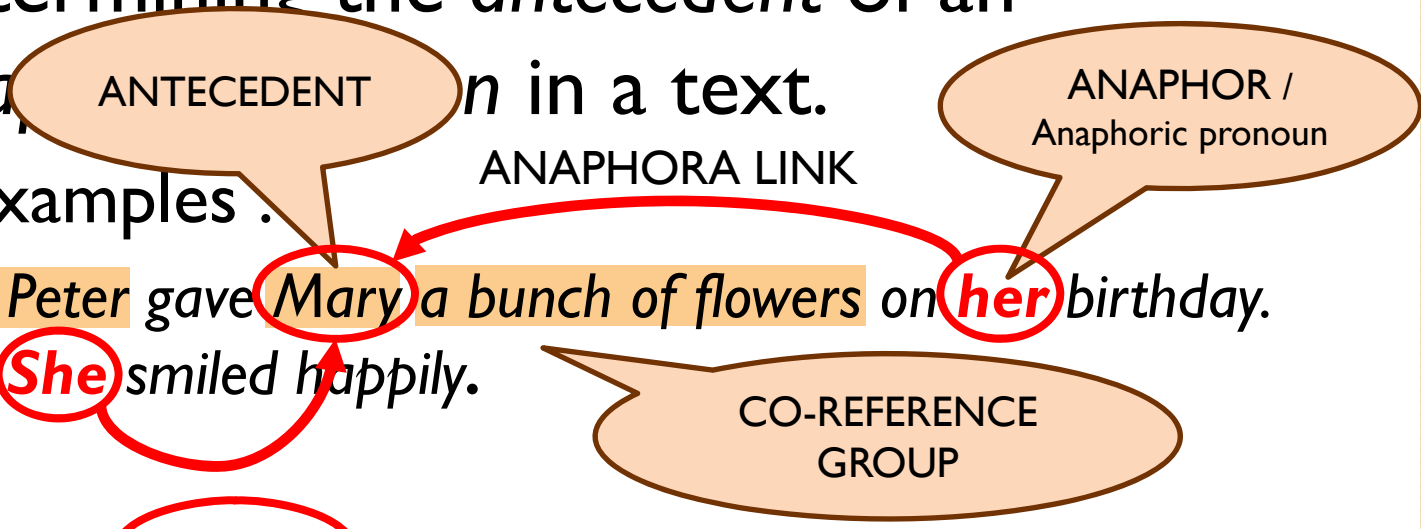
LREC 2008

Objective

- What are the difficulties in PR for the biomedical domain compared with other domains ?
- What kinds of features are useful for the biomedical domain ?

Pronoun resolution (PR)

- ***Pronoun resolution*** : the NLP task of determining the *antecedent* of an *anaphoric pronoun* in a text.
  - Examples :
    - *Peter gave Mary a bunch of flowers on her birthday. She smiled happily.*



    - *The IL-2 gene displays both T cell specific and inducible expression. It is only expressed in CD4+ T cells after antigenic or mitogenic stimulation.*

What we did

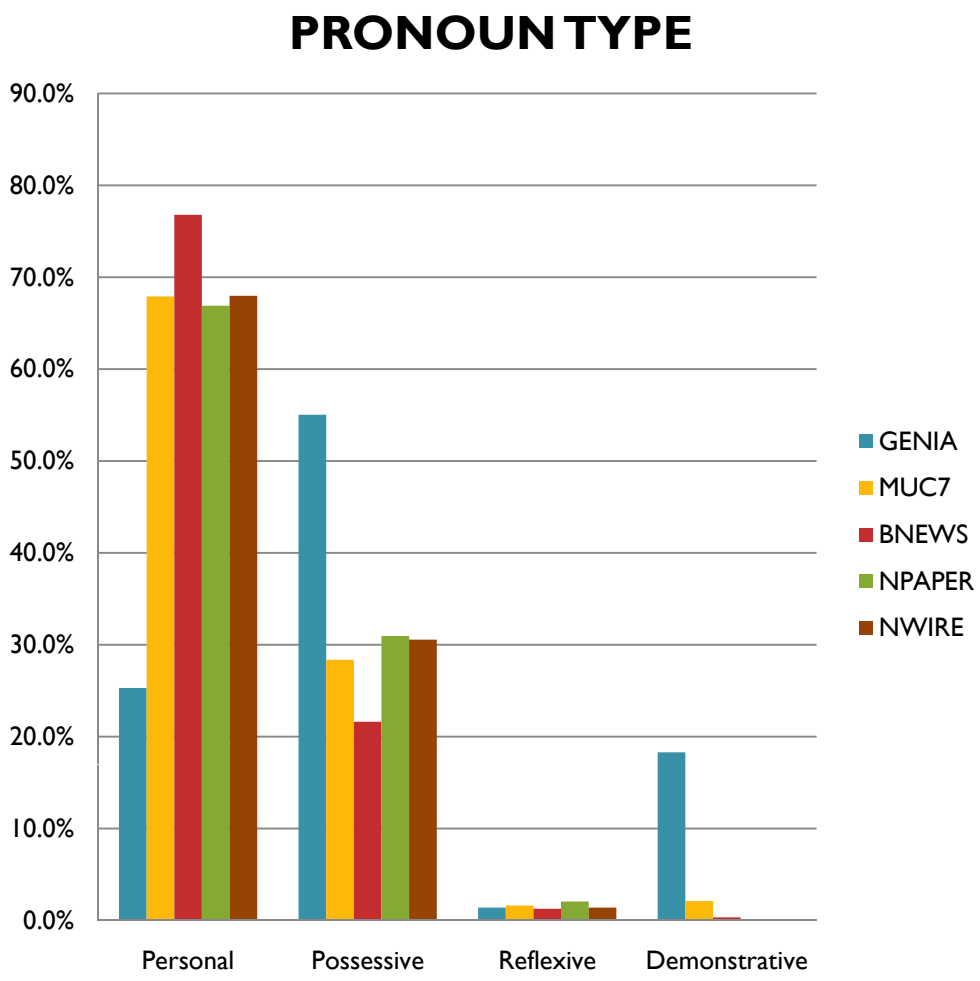
- Analyzing the differences of 3 corpora
  - MUC and ACE for the news wire domain
  - GENIA for the bio-domain
- Building a machine-learning based pronoun resolution system
- Comparing the ***contributions of features*** for each corpus

Corpora

- Data sets

Data set	Training set (no. of anaphoric pronouns)	Size of test set (no. of anaphoric pronouns)
GENIA	1442	357
ACE-BNEWS	2427	633
ACE-NPAPER	2058	613
ACE-NWIRE	2177	450
MUC-7	371	240

Corpus analysis (Pronoun type)

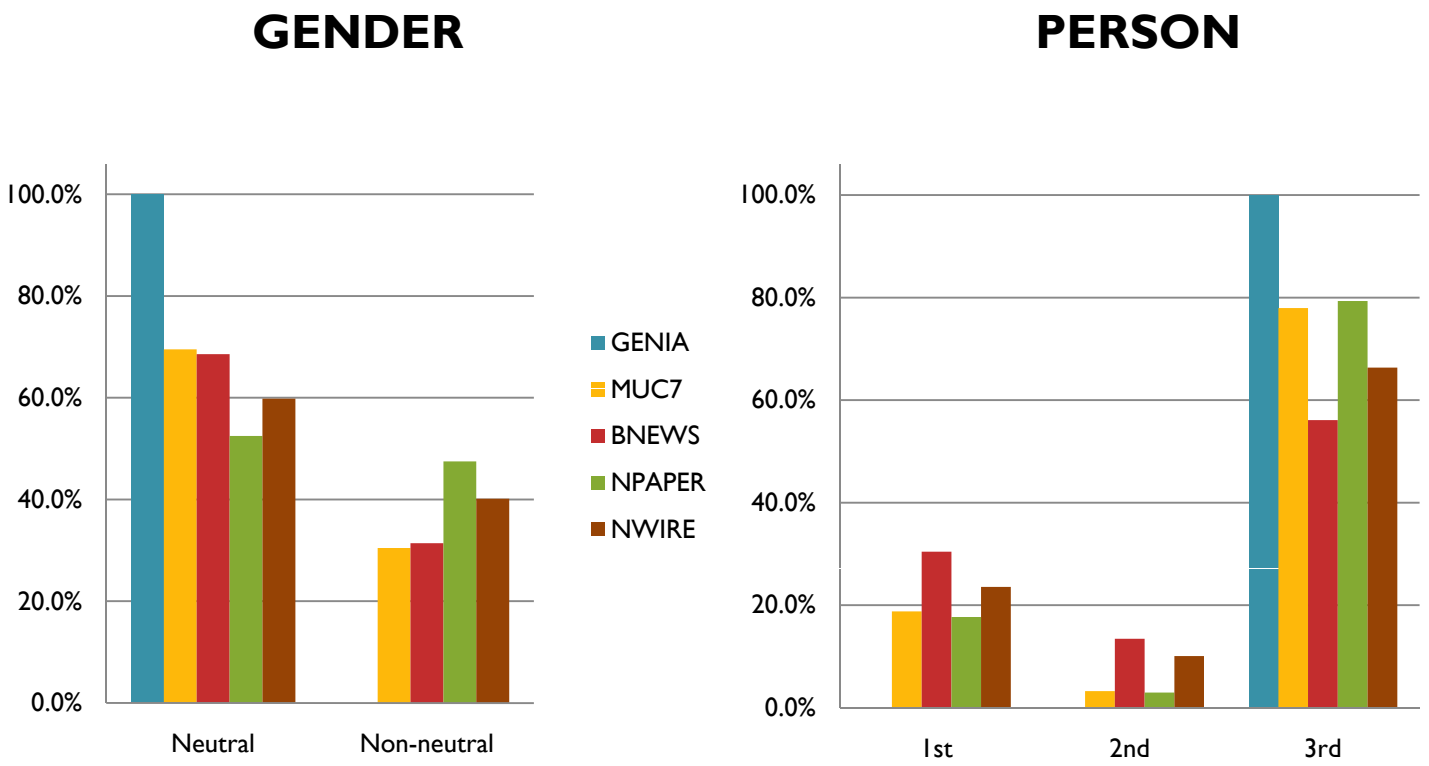


GENIA contains

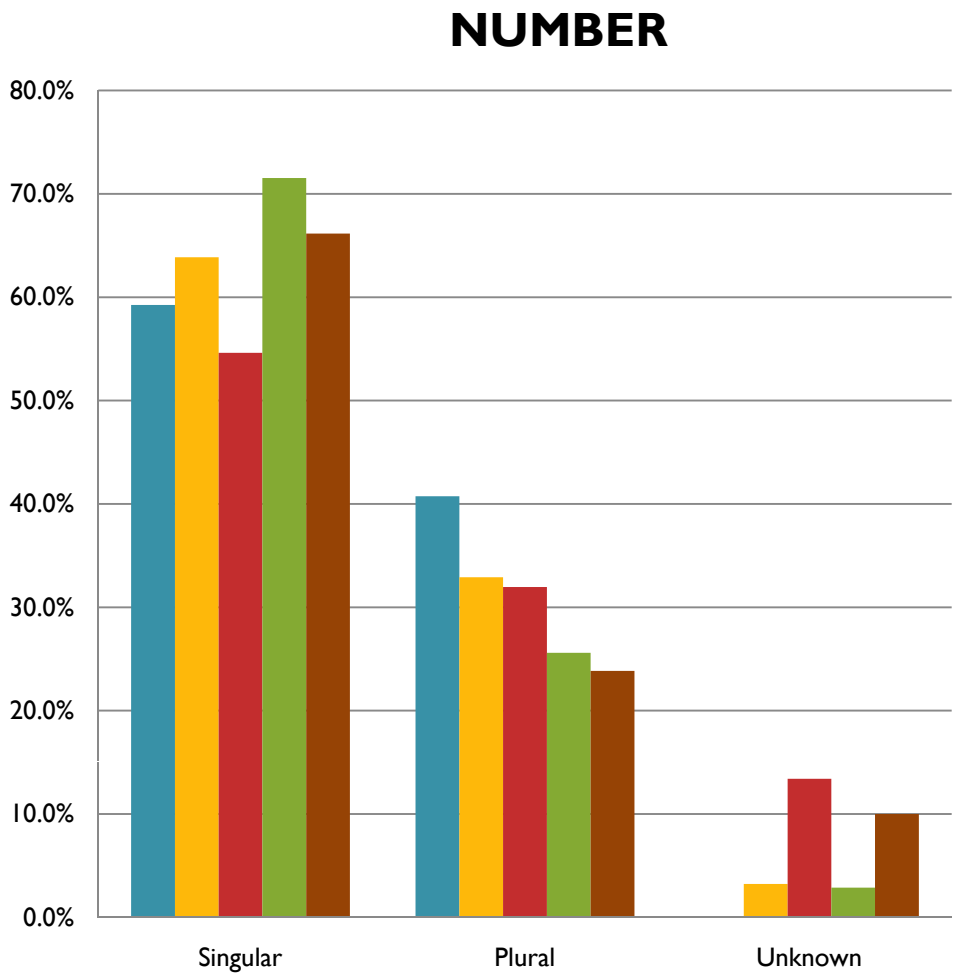
- More ***demonstrative*** (e.g. this, those) and ***possessive*** pronouns (e.g. its, their).

Corpus analysis (Gender, Person)

- All of the anaphoric pronouns in GENIA are ***neutral-gender*** *and* ***third-person***.



Corpus analysis (Number)



GENIA:

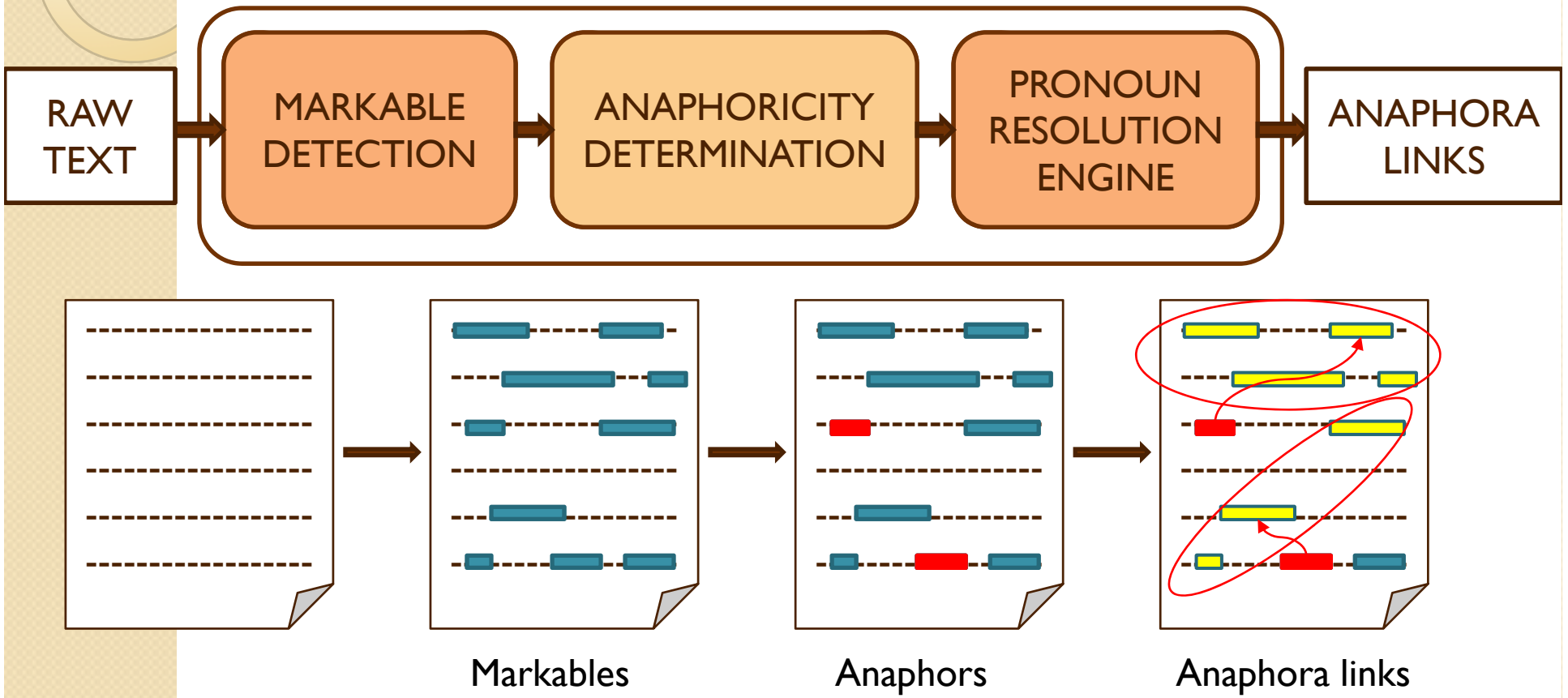
- All are ***number-deterministic*** (i.e. either singular or plural)

Summary of corpus analysis

- Compared to the other corpora, in the GENIA corpus:
  - There are many *demonstrative and possessive* pronouns.
  - All of the pronouns *are neutral-gender and third-person*.
  - All of the pronouns are *number-deterministic*.

- → Except for the *number* property, pronouns in bio-texts contain very ***poor information*** about their antecedents.

Pronoun resolution system



Experiment with Markable detection

- Using a POS tagger and a base-NP chunker
- Results

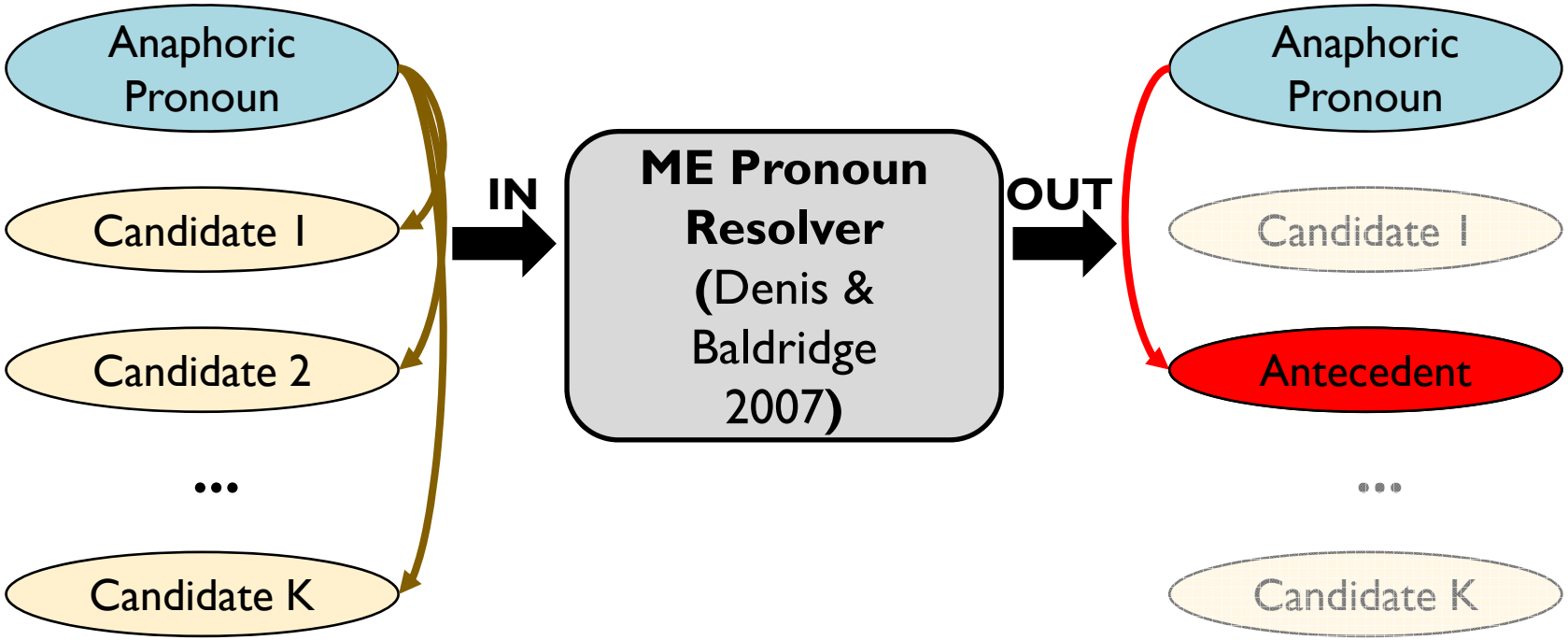
	GENIA	ACE	MUC
Mention coverage	94.59%	95.66%	94.46%
Link coverage	89.55%	92.98%	90.76%

- Difficulties in the biomedical domain
  - Complex markables (e.g., *1 1 alpha-methyl-1alpha,25-(OH)2D3*)
  - Coordinated markables
  - POS-ambiguous *that*

Experiments with Pronoun resolution engine

- Is the main component of the system, containing the PR model

Pronoun resolution engine

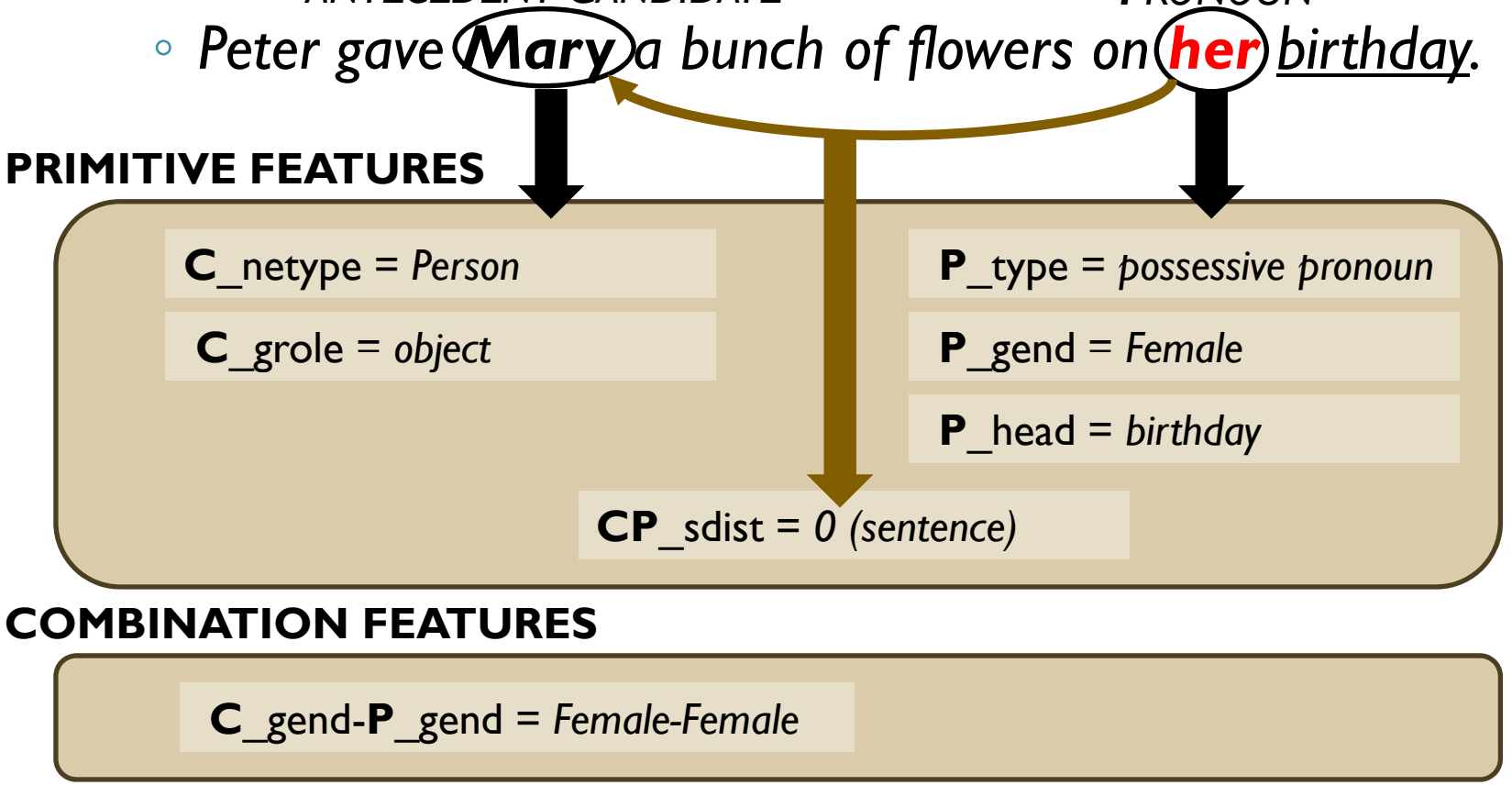


- Using gold mentions annotated in the corpora.
- Evaluation score : (Mitkov 2001)

$$Success\ rate = \frac{\text{Number of successfully resolved anaphors}}{\text{Number of all anaphors}}$$

Features

- An input link is characterized with the *combinations* of 25 *primitive features*.
  - *Peter gave* **Mary** *a bunch of flowers on* **her** *birthday*.



Feature groups

FEATURE SET	FEATURE GROUP	PRIMITIVE FEATURE	INFORMATION
Fundamental	mention type	P_type C_type	Morphological
Baseline	**sdist**	CP_sdis	
	tdist	CP_tdis	
	numb	P_numb C_numb	
	pers	P_pers C_pers	
	gend	P_gend C_gend	
	pfam	P_pfam C_pfam	
	string	P_word C_head	
Additional	**pos**	P_lpos P_rpos C_lpos C_rpos	Syntactic
	grole	P_grole C_grole	
	netype	C_netype	Semantic
	last3c	C_last3c	Morphological
	comb	P_head	Syntactic
		C_1stnp	Discourse

Experiment 1: Contributions of the baseline features

Excluded(-)	GENIA (bio)	ACE (nw)	MUC (nw)
BASELINE	**70.31**	**64.61**	**57.08**
-sdist	67.23(-3.08)	63.51(-1.10)	51.67(-5.41)
-tdist	70.03(-0.28)	59.56(-5.05)	57.08(+0.00)
-numb	65.83(-4.48)	61.77(-2.84)	58.33(+1.25)
-pers	70.31(+0.00)	57.19(-7.42)	55.42(-1.66)
-gend	69.75(-0.56)	64.45(-0.16)	56.67(-0.41)
…			

- ***numb*:**
  - is *the most effective feature* in the bio-domain.
  - →GENIA: pronouns are ***number deterministic***

Experiment 2: Contributions of the additional features

Included (+)	GENIA (bio)	ACE (nw)	MUC (nw)
BASELINE	**70.31**	**64.61**	**57.08**
+pos	75.63(+5.32)	62.88(-1.73)	57.50(+0.42)
+grole	73.67(+3.36)	63.82(-0.79)	58.75(+1.67)
+netype	73.95(+3.64)	64.30(-0.31)	58.33(+1.25)
…			

- ***netype: C_netypeP_head***
  - *Tax is thought to be crucial in the development of the disease, since it transforms healthy T cells in vitro and induces tumors in transgenic animals.*

Semantic preference: (PROTEIN , transform)

Experiment 2: Contributions of the additional features (cont)

Included (+)	GENIA	ACE	MUC
BASELINE	**70.31**	**64.61**	**57.08**
+pos	75.63(+5.32)	62.88(-1.73)	57.50(+0.42)
+grole	73.67(+3.36)	63.82(-0.79)	58.75(+1.67)
+netype	73.95(+3.64)	64.30(-0.31)	58.33(+1.25)
…			

- ***grole: CA_sdist-C_parg-P_parg***
  - ***Fludarabine*** *is a nucleoside analog used in the treatment of hematologic malignancies that can induce severe and prolonged immunosuppression . Although* ***it*** *can be incorporated into the DNA of …*

SUBJECTS !

Integration of the additional features

- Integrating all positive features for each corpus resulted in:
  - the significant increase in the success rate of GENIA (bio-domain)

	GENIA	**ACE**	**MUC**
BASELINE	**70.31**	**64.61**	**57.08**
INTEGRATION	79.55 (+9.24)	64.61 (+0.00)	60.42 (+3.34)

Conclusion

- Entity mentions in the biomedical domain are complex, which makes it difficult to extract markables.
- Anaphoric pronouns in bio-texts contain very poor information about their antecedents.
- Context information of the pronouns plays a very important role in the PR.

Future work

- Improving the markable detection component
- Exploiting more syntactic features
- Integrating the pronoun resolution into an information extraction system