Extracting Protein Interactions from Text with the Unified AkaneRE Event Extraction System

Rune Sætre, Kazuhiro Yoshida, Makoto Miwa, Takuya Matsuzaki, Yoshinobu Kano, and Jun’ichi Tsujii

IEEE/ACM Transaction on Computational Biology and Bioinformatics, 2010

(Presentation by Michael Gabilondo)

Introduction

- AkaneRE is a configurable system to learn from different annotated training data (BioNLP, BioCreative, REMerge, AIMed are current supported formats)
- It can predict binary, undirected PPI (AIMed training data), or complex events with nested sub-events, and variable number of arguments and semantic roles (GENIA/BioNLP training data)
- It the best result on the REMerge corpora, but ranks 6th in BioNLP, and among the top 3 in BC IPT and BC INT tasks

Outline

- **Background and shared tasks**
  - Enju
  - Brief summary of AkanePPI paper (2008)
  - BioNLP-EE and BioCreative II.5 tasks
- **Brief summary of BioNLP-EE paper (2009)**
  - AkanePPI becomes AkaneRE
- **Unified AkaneRE system (2010, this paper)**
  - BioCreative II.5 system

Part 1

Background and Shared Tasks

Enju

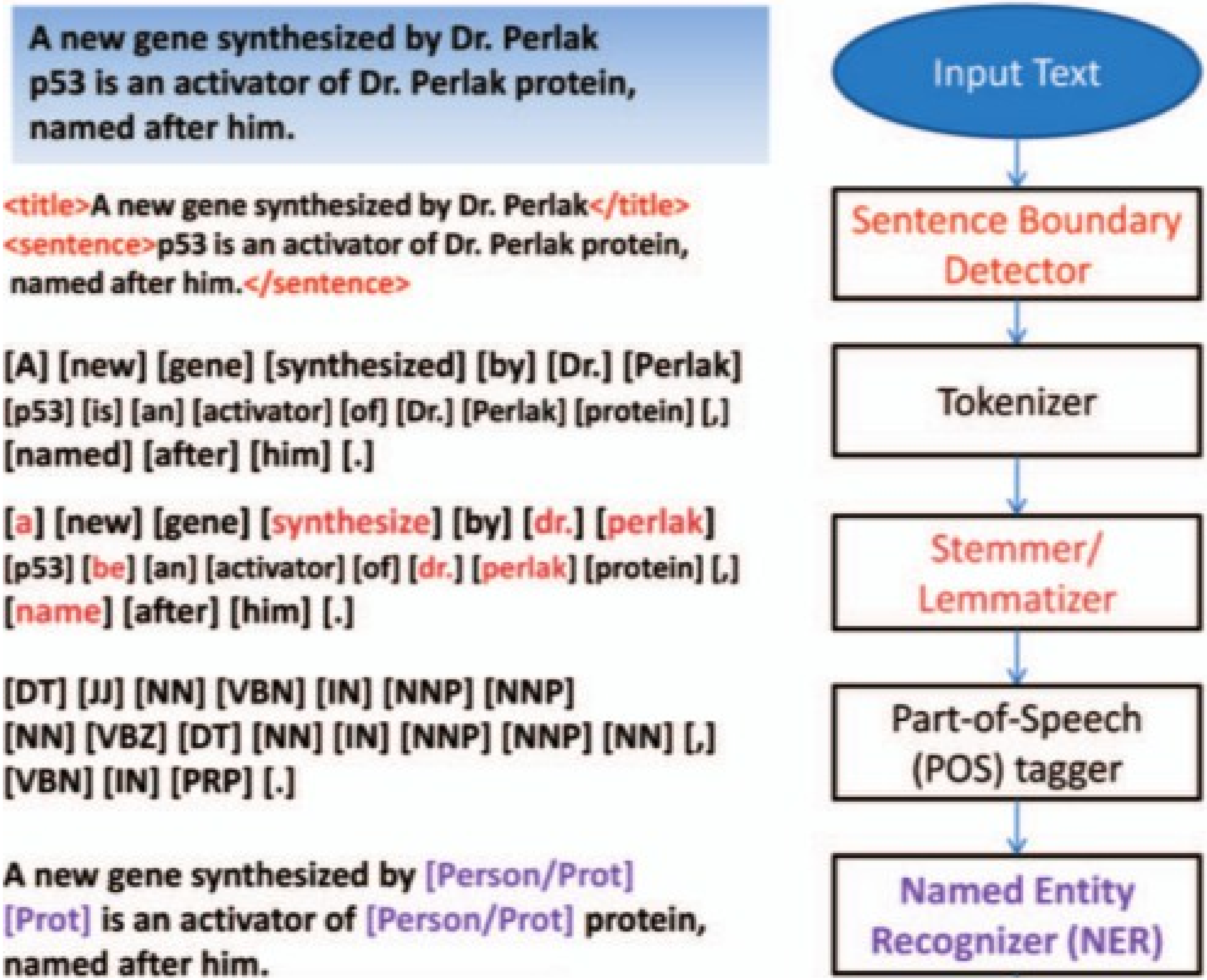
- Enju is a Head-driven Phrase Structure Grammar (HPSG) parser
- Outputs both phrase structures and predicate-argument structures
- Every word is a predicate, and there are many predicate types, so the argx are interpreted differently depending on the predicate type
- Includes parsing model for biomedical domain

Enju

- ``John has come''
- Phrase structure: (S (NP John) (VP has (VP come)))
- PAS: <predicate, relation, argument phrase/clause>
  - < come, arg1, John > < has, arg1, John > < has, arg2, come >
  - arg1 is the ``semantic subject'', so takes passive constructions into account

AkanePPI System

- AkanePPI initially trained on AIMed corpus, as described in “Syntactic features for protein-protein interaction extraction” (Sætre, Sagae, Tsujii, 2008)
- PAS paths between protein pairs, generated by Enju; features represented as trees
- Also used features from GDep (GENIA Dependency parser), and BOW for before/after/in-between proteins
- Used SVM-light with Tree-Kernels



Input Text

A new gene synthesized by Dr. Perlak
p53 is an activator of Dr. Perlak protein,
named after him.

Sentence Boundary Detector

<title>A new gene synthesized by Dr. Perlak</title>
<sentence>p53 is an activator of Dr. Perlak protein,
named after him.</sentence>

Tokenizer

[A] [new] [gene] [synthesized] [by] [Dr.] [Perlak]
[p53] [is] [an] [activator] [of] [Dr.] [Perlak] [protein] [,]
[named] [after] [him] [.]

Stemmer/ Lemmatizer

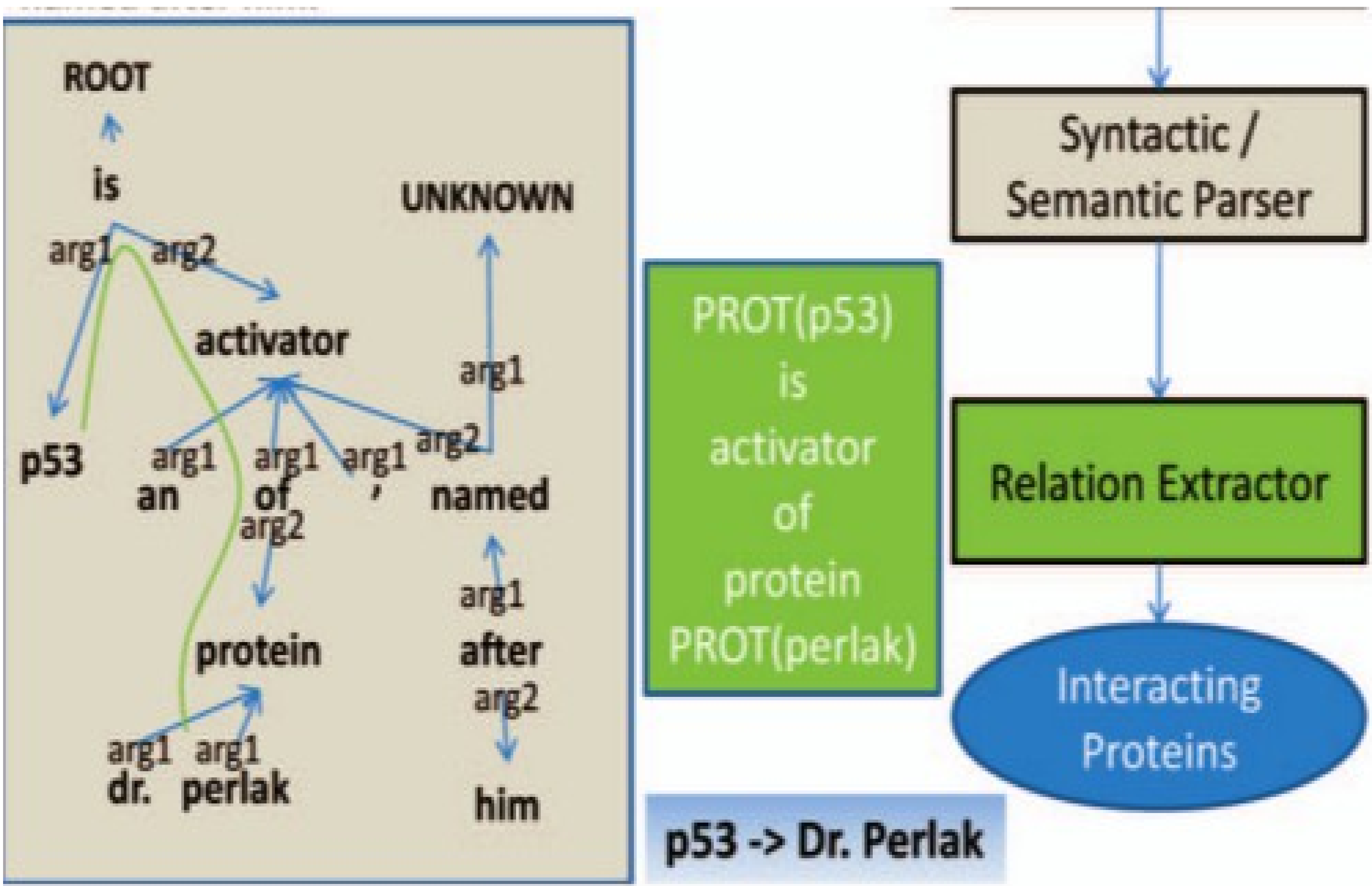
[a] [new] [gene] [synthesize] [by] [dr.] [perlak]
[p53] [be] [an] [activator] [of] [dr.] [perlak] [protein] [,]
[name] [after] [him] [.]

Part-of-Speech (POS) tagger

[DT] [JJ] [NN] [VBN] [IN] [NNP] [NNP]
[NN] [VBZ] [DT] [NN] [IN] [NNP] [NNP] [NN] [,]
[VBN] [IN] [PRP] [.]

Named Entity Recognizer (NER)

A new gene synthesized by [Person/Prot]
[Prot] is an activator of [Person/Prot] protein,
named after him.



Syntactic / Semantic Parser

PROT(p53) is activator of protein PROT(perlak)

Relation Extractor

Interacting Proteins

p53 -> Dr. Perlak

Results from AkanePPI paper

- 10-fold cross validation
- Best precision using PAS Enju features alone
  - 72% precision, 28.7% recall, 41% F-score
- Best Recall and best F-score using Gdep, Enju and BOW features
  - 64.3% precision, 44.1% recall, 52% F-score
  - Many percent-points higher than state-of-the-art (2008)

BioNLP-EE (shared task #1)

- Extract bio-events on proteins or genes (which are not distinguished)
  - There are 9 event types
    - Binding, Gene_expression, Localization, Negative_regulation, Phosphorylation, Positive_regulation, Protein_catabolism, Regulation, and Transcription
- It is assumed NER has been performed, and the gold standard provides all protein/genes in the text

BioNLP-EE (shared task #1)

- Number of entities/arguments in the event can more than 2
- Agent and Theme are distinguished, so the relationships are not symmetric
- 800 abstracts, training data
- ***Side note****:* Task 2 involves finding secondary arguments, with different semantic roles, and heads other than protein/genes.
    - In GENIA, other entities than proteins ARE annotated (they have a term and event ontology), but roles other than theme and cause are not

BioCreative II.5 (training data)

- Training data is 740 full-text articles
- Extracting binary, undirected relationships between proteins
- Gold standard annotation is given at the article-level
    - For a given article, we know only the Uniprot accession numbers of the proteins in the article
    - And we know the pairs of proteins that interact

BioCreative II.5 (INT and IPT tasks)

- Interaction normalization task (INT)
  - Report the protein accession numbers in the testing data
  - **Each of the N hits is ranked with a unique ID in [1...N]**
  - also provide a confidence value, which is used to break ties but not in calculation of performance measure
- Interaction Pair Task requires INT as a subtask
  - Report pairs of accession numbers, and the pairs are ranked

BioCreative II.5 (AUC)

- Since the output is ranked, BC uses AUC as its performance measure
- AUC := Area under the interpolated P/R curve
  - P/R graph has Recall on x-axis and Precision on y-axis; higher recall levels correspond to lower ranked results
  - P/R curve is jaggy, and the interpolated P/R smooths it
    - i.e., retrieving the next ranked document causes sharp increases or decreases in precision or recall
- *Optimizing for AUC leads to low F-scores, so a cut-off in rank/confidence must be selected*

REMerge PPI Corpora

- There also exist five PPI corpora, annotated with proteins and interactions at the text level (not article level)
  - AIMed
    - 177 abstracts with interactions, 48 abstracts without interactions (interactions may exist, just not at sentence level, and not annotated)
  - BioInfer, HPRD50, IEPA, LLL
- Popular corpora for PPI evaluation

Part 2

Extending AkanePPI
to handle
Bio-Events in GENIA Event Corpus
(BioNLP-EE task)

Brief summary of 2009 paper

BioNLP-EE system

Tokenization, POS tagging, Parsing (Enju & Gdep)

- Event Clueword Recognition
    - GENIA annotates the clueword triggering each event, e.g., a verb
    - Used NER system to tag clueword with one of the 8 (9?) event types
- Event Template Extraction
    - Extracted 9 generalized templates, which contain syntactic and semantic information about arguments (# of arguments, semantic roles, NE type of each role)
- Learn which ne/event combinations go in each template, based on training data

Freq	**Event**	**Theme1**	**Theme2**	**Theme3**	**Theme4**	**Cause**
-	PPI	Protein	Protein			
613	Binding	Protein				
213	Binding	Protein	Protein			
3	Binding	Protein	Protein	Protein		
2	Binding	Protein	Protein	Protein	Protein	
217	Regulation	Protein				Protein
12	Regulation	Binding				Protein
48	+Regulation	Transcription				Protein
4	+Regulation	Phosphorylation				Binding
5	-Regulation	+Regulation				Protein
...	...	...				...
Total	148 Templates					
Count	General Templates	**Theme1**	**Theme2**	**Theme3**	**Theme4**	**Cause**
9	event templates	Protein				
1	event template	Protein	Protein			
1	event template	Protein	Protein	Protein		
1	event template	Protein	Protein	Protein	Protein	
3	event templates	Protein				Protein
12	event templates	Protein				Event
27	event templates	Event				
26	event templates	Event				Protein
68	event templates	Event				Event

Learning template instances

- Each generalized template was matched with all legal combinations of named entities, including event classes (clue words) and proteins
- One logistic regression classifier (LIBLINEAR) was learned for each generalized template class, using one-vs-rest
    - Features include dependency paths, BOW
- Each template instance is mapped to a confidence value, and they chose a cut-off threshold by hand
- If a highly confident regulation event includes sub-events that are below the confidence threshold, those sub-events are *still* output; this was changed in newer paper, so that all sub-events must be about threshold

Results

- **Template classes do not correspond to event classes; clue word detection is necessary, and it has low accuracy (50%) which was their major shortcoming**
  - ``closer integration between clue-word recognition and template prediction modules can lead to better performance''
- Official F-score is 36.9%, came in 6th out of 24 groups
- Allowing the system to predict multiple confident alternatives for the same event-word raised F-score to 42.6%
  - One clueword can belong to two different event categories

Part 3

Unified AkaneRE System (2010)

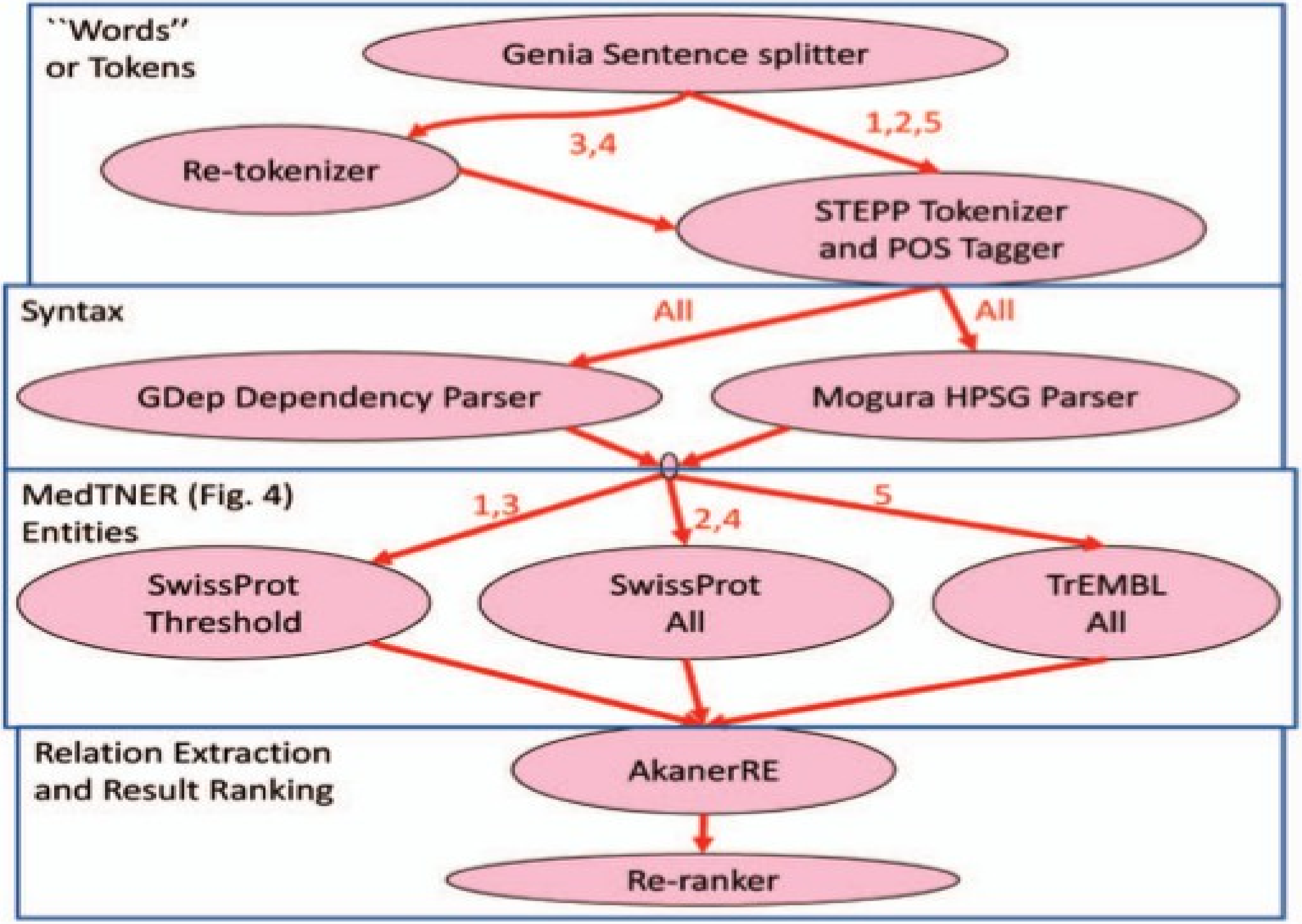
Biocreative II.5 IPT and INT tasks

(Extending and applying the 2009 system for the BC II.5 tasks)

Unified AkaneRE

- This system uses a configuration file that allows the user to specify, for example,
  - Learning/prediction/cross-validation mode
  - Gold/training data; this file determines the kinds of predictions the system handles, either PPI or event
  - (Optional) Information about the NER system
  - List of features from parser output
  - Parameters for machine learning system
- Everything uses (or must be converted to) Standoff annotation, an external annotation (and not inline), where the annotation file contains pointers into the original text file
  - Many tools can easily add small annotations to the text file
- UIMA, a framework for standard way of annotationg and sharing information about free text

BC II.5 System: BCMS Interface and U-Compare workflows



Named Entity Recognition

- MedTNER identifies proteins in text and maps them to Uniprot accession numbers
- Dictionary lookup does string matching and annotates matches with synonym lists from Uniprot, Entrez Gene, GENA dictionaries
- False positive filtering uses a binary classifier (logistic regression) trained on GENIA to annotate matches with confidence scores
- From the above, the system knows which are proteins and which are not, and next, it maps those that are to Uniprot accession numbers in the disambiguation module

NER: Disambiguation Module

- The disambiguation module is trained on the BC corpus, which has article-level annotations, so the features are also article level
  - It learns the kinds of articles the NE can appear in
  - Features include similarity between target document and all documents linked to each dictionary entry

Interaction Detection Module

- The interaction module was trained on AIMed, since that has text-level annotated proteins and interactions
- “It puts together all combinations of entities into events and assigns a confidence score saying whether that combination is likely or unlikely”
- This probably uses similar PAS features as 2008 paper, but they have switched from Tree-Kernels to Logistic Regression using LIBLINEAR, which assigns a confidence score from 0 to 1 for all predictions

Reranking and iPR-AUC optimization

- Rerank the pairs predicted by the Akane interaction detection module
    - Those pairs are using only AIMed data on BC task
    - Rerank to take into account BC article-level training data
    - Use article level features
- To create features, the sentence is enriched with species information from NCBI Entrez Taxonomy
- Uses logistic regression

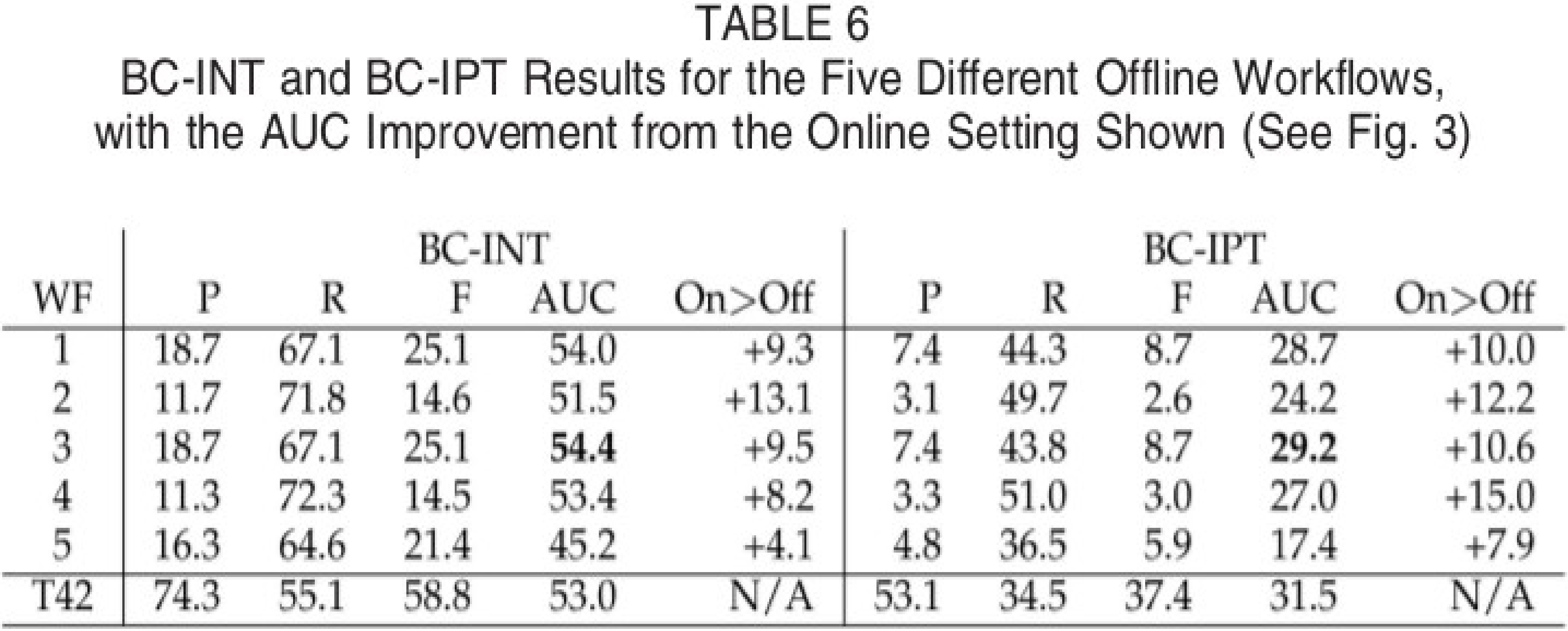
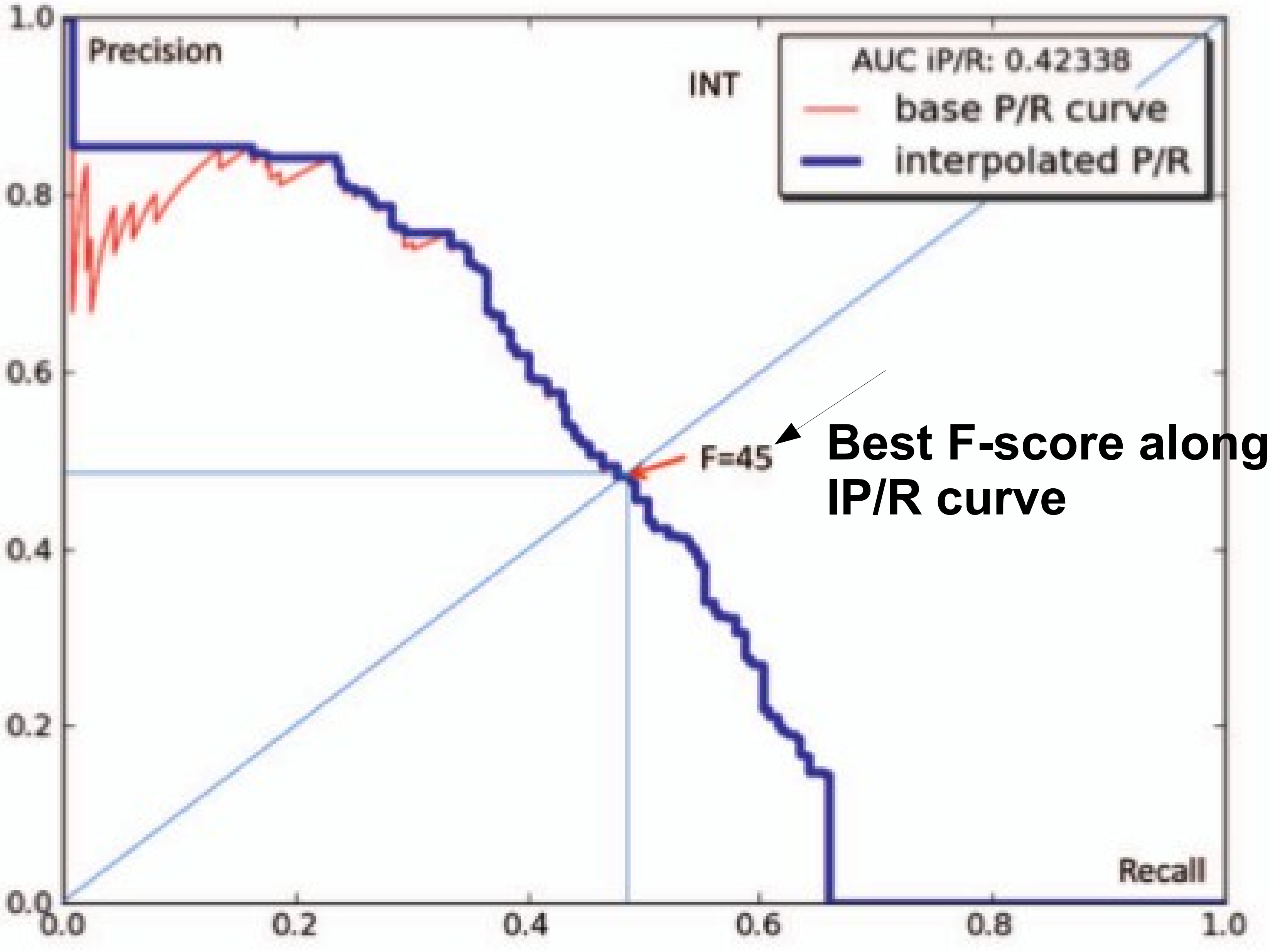


TABLE 6
BC-INT and BC-IPT Results for the Five Different Offline Workflows, with the AUC Improvement from the Online Setting Shown (See Fig. 3)

	BC-INT					BC-IPT				
WF	P	R	F	AUC	On>Off	P	R	F	AUC	On>Off
1	18.7	67.1	25.1	54.0	+9.3	7.4	44.3	8.7	28.7	+10.0
2	11.7	71.8	14.6	51.5	+13.1	3.1	49.7	2.6	24.2	+12.2
3	18.7	67.1	25.1	**54.4**	+9.5	7.4	43.8	8.7	**29.2**	+10.6
4	11.3	72.3	14.5	53.4	+8.2	3.3	51.0	3.0	27.0	+15.0
5	16.3	64.6	21.4	45.2	+4.1	4.8	36.5	5.9	17.4	+7.9
T42	74.3	55.1	58.8	53.0	N/A	53.1	34.5	37.4	31.5	N/A

- BC task was to optimize for AUC
- They could provide all results, ranked and did not need a cut-off threshold, to optimize AUC
- As a result, precision suffers, since precision does not take into account rank, and considers the data as unordered.

  P = TP / (TP + FP) ~ rank is not taken into account, so many FP
- Need to choose a cut-off Threshold to improve precision



Best F-score along IP/R curve

g. 5. Biocreative protein INT results.

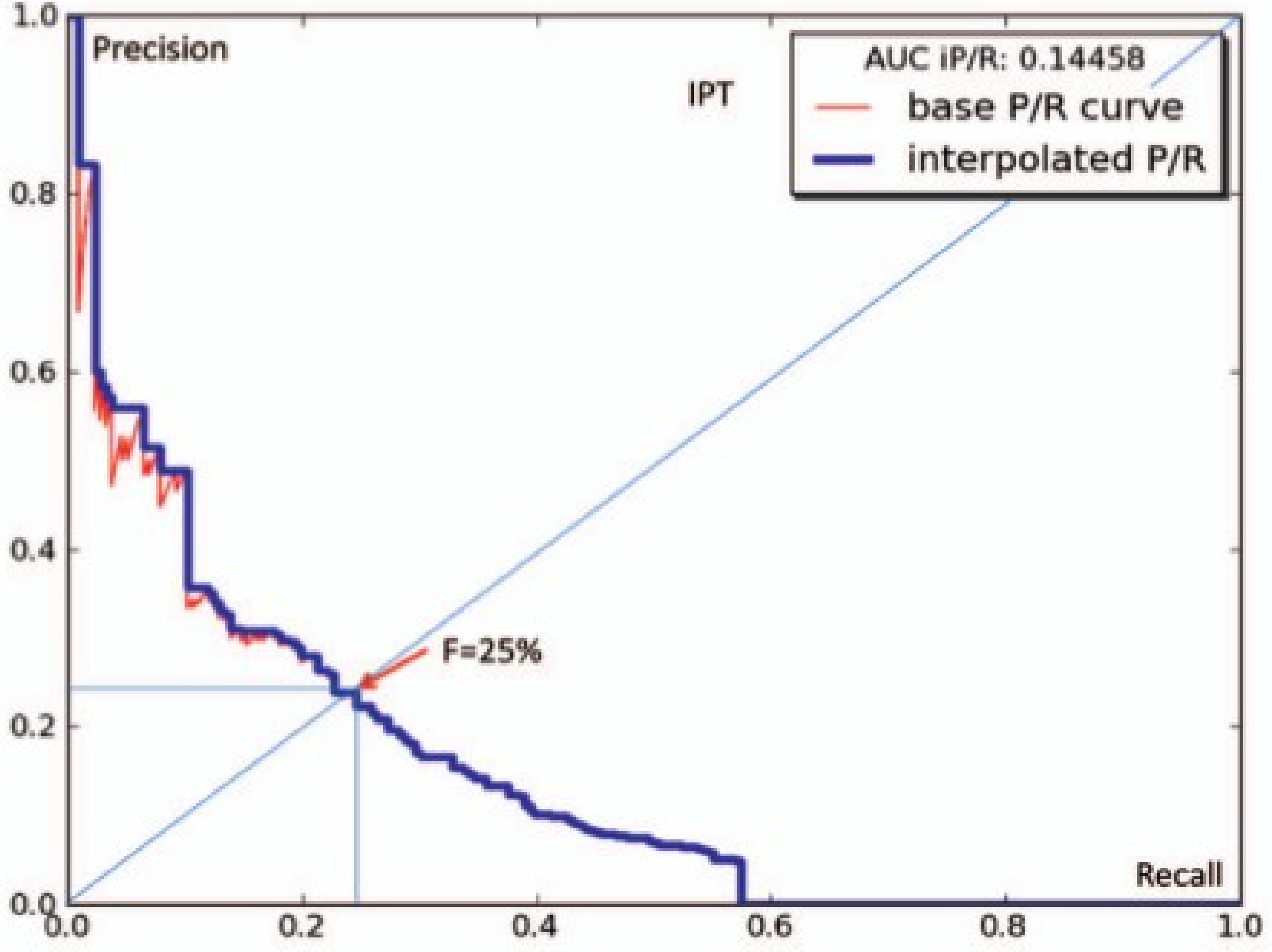


Fig. 6. BC-IPT results.

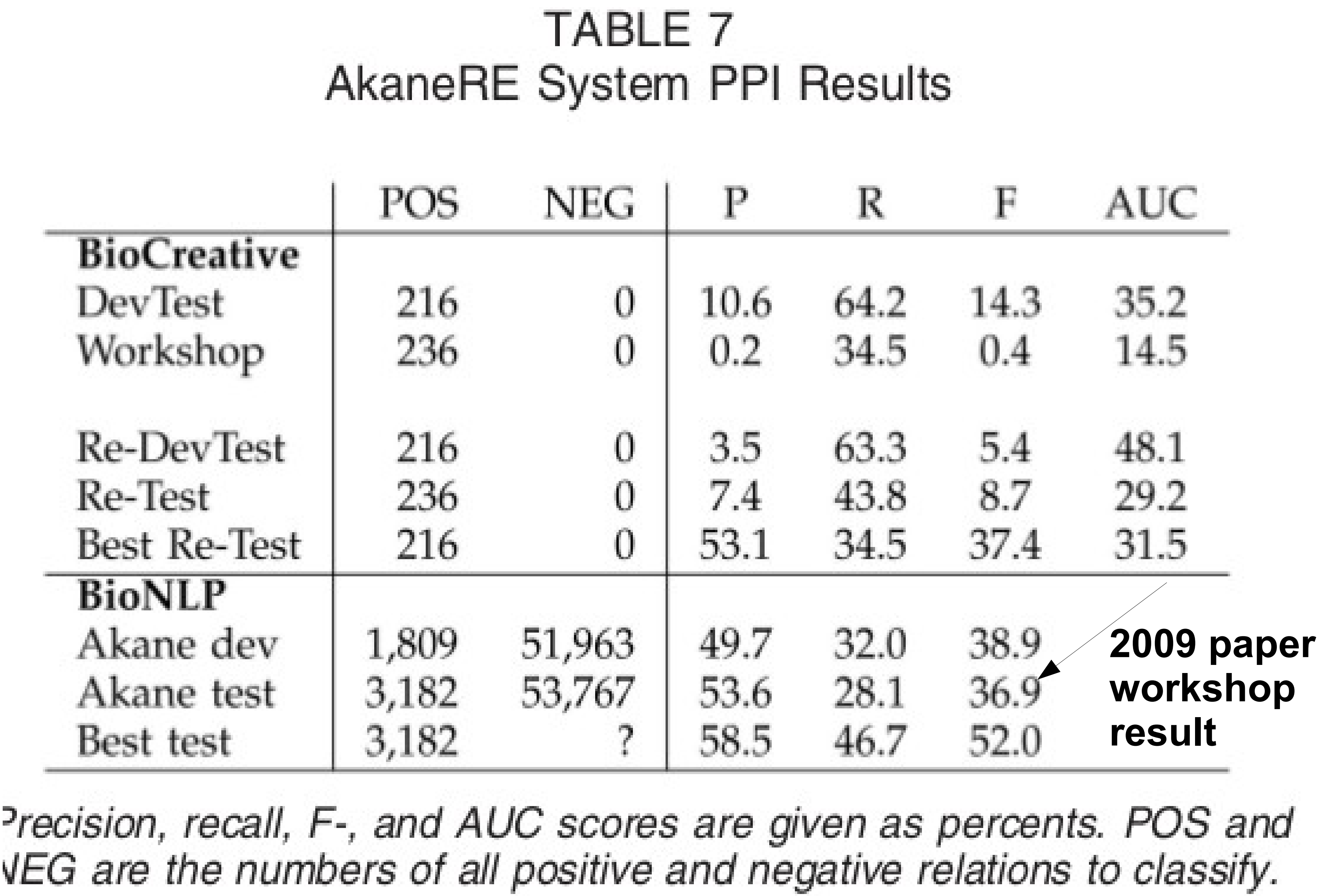


TABLE 7
AkaneRE System PPI Results

	POS	NEG	P	R	F	AUC
BioCreative						
DevTest	216	0	10.6	64.2	14.3	35.2
Workshop	236	0	0.2	34.5	0.4	14.5
Re-DevTest	216	0	3.5	63.3	5.4	48.1
Re-Test	236	0	7.4	43.8	8.7	29.2
Best Re-Test	216	0	53.1	34.5	37.4	31.5
BioNLP						
Akane dev	1,809	51,963	49.7	32.0	38.9	
Akane test	3,182	53,767	53.6	28.1	36.9	
Best test	3,182	?	58.5	46.7	52.0	

Precision, recall, F-, and AUC scores are given as percents. POS and NEG are the numbers of all positive and negative relations to classify.

2009 paper workshop result

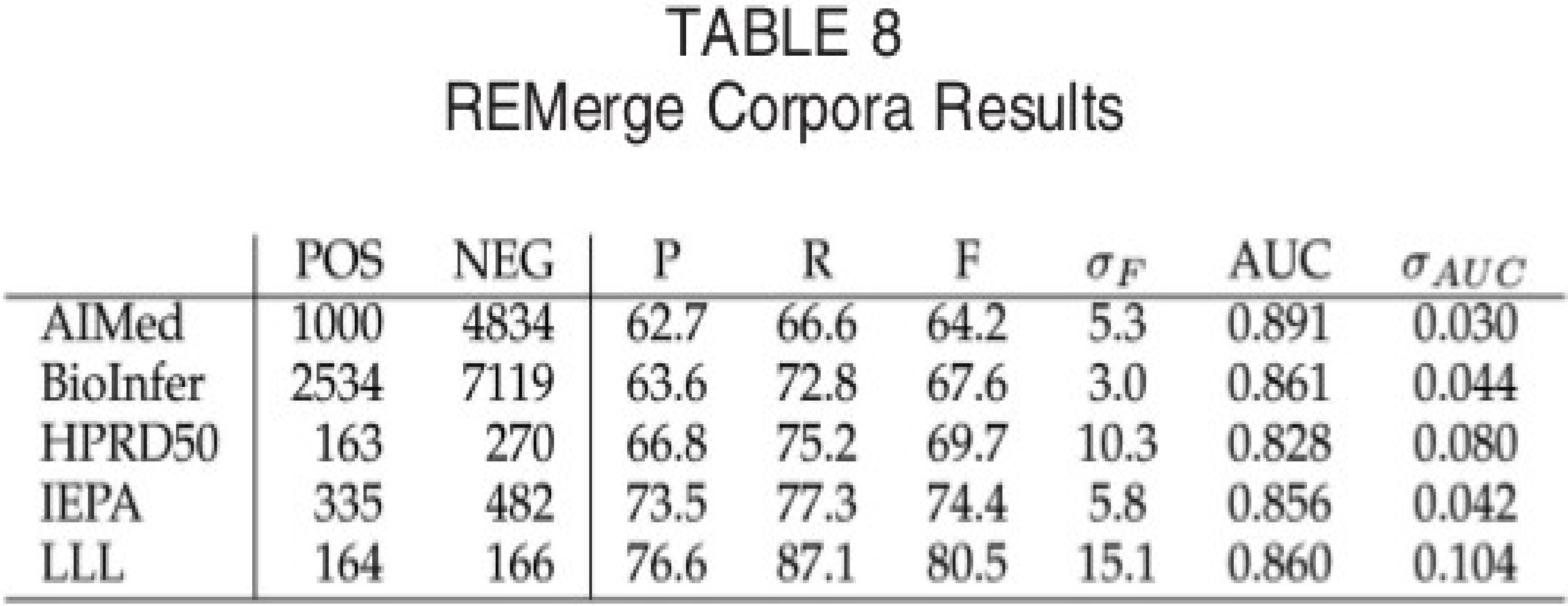


TABLE 8
REMerge Corpora Results

	POS	NEG	P	R	F	σ_F	AUC	σ_{AUC}
AIMed	1000	4834	62.7	66.6	64.2	5.3	0.891	0.030
BioInfer	2534	7119	63.6	72.8	67.6	3.0	0.861	0.044
HPRD50	163	270	66.8	75.2	69.7	10.3	0.828	0.080
IEPA	335	482	73.5	77.3	74.4	5.8	0.856	0.042
LLL	164	166	76.6	87.1	80.5	15.1	0.860	0.104

- Best F-Score in 2008 paper was 52% (AIMed).
- Here, AIMed has 64.2% F-score

Appendix: Logistic Regression

Logistic regression is regression for binary targets.

The model learns the regression coefficients (bi),

$z = b0 + b1x1 + b2x2 + .. + bnxn$,

where z is the target variable, and the xi are the independent features/variables (This is like regression for continuous targets).

The z is continuous from minus infinity to positive infinity, **and the logistic function maps this range to the range (0,1)**.

This may be interpreted as a confidence value; the AkaneRE system does not choose a cut-off to optimize AUC, but has negative results for F-score.

Logistic function is 1 / (1 + e^{-z})