

FactRuEval-2016: Information Extraction based on deep syntactic-semantic analysis

Stepanova M. E., Budnikov E. A., Chelombeeva A. N., Matavina P. V., Skorinkin D. A.

Dialogue 2016

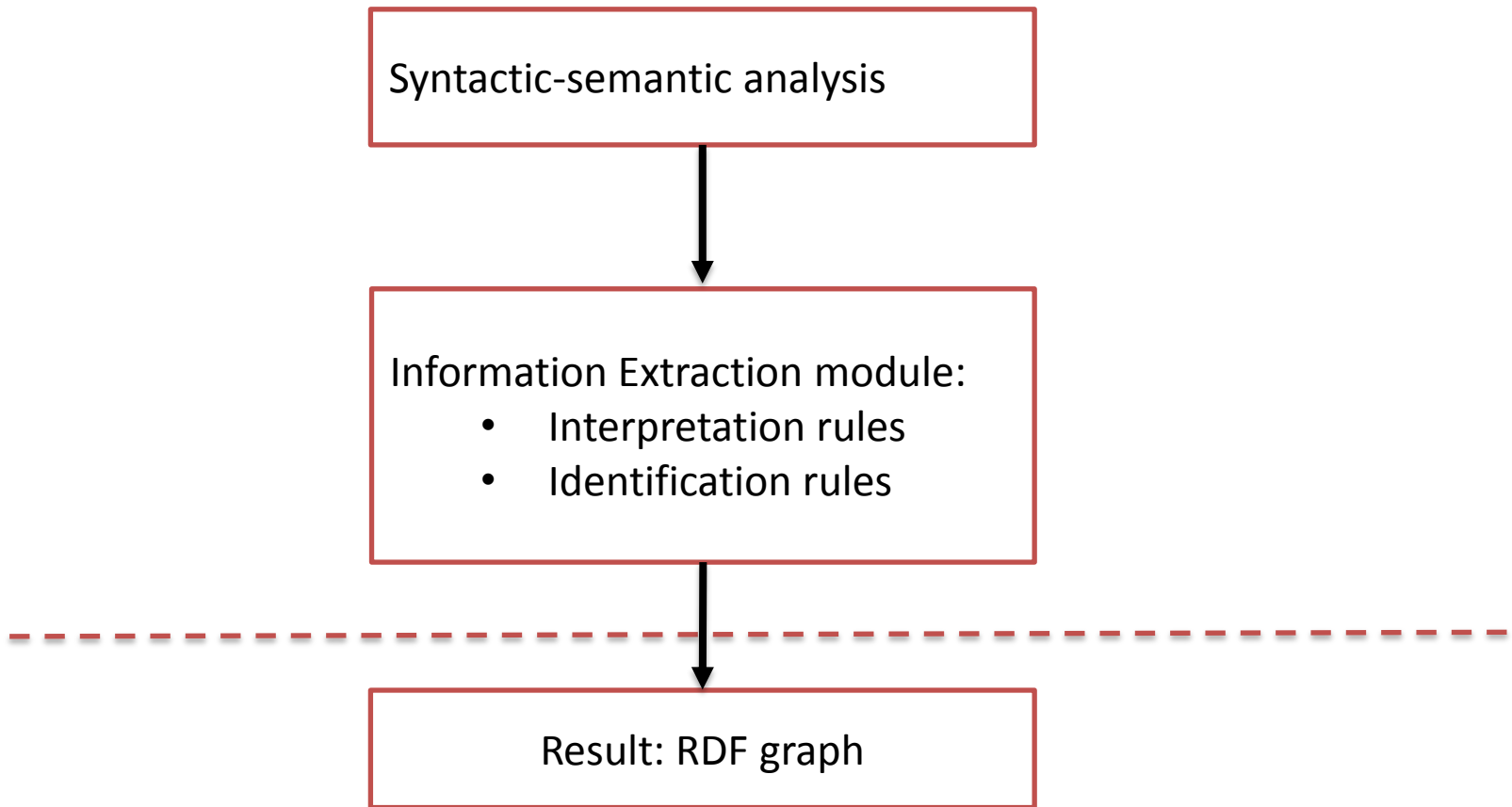
Outline

- FactRuEval overview
- Information Extraction Mechanism
- Results of the track

FactRuEval Overview

- Track 1. Named Entity Recognition
 - Comparison with Location/LocOrg differentiation
 - Comparison without Location/LocOrg differentiation
- Track 2. Named Entity Recognition with attributes and local identification
- Track 3. Fact Extraction

Information Extraction Mechanism. Overview



Information Extraction Mechanism.

Место спикера Палаты представителей в новом Конгрессе занял республиканец Джон Бенер.

Information Extraction Mechanism.

Место спикера **Палаты представителей** в новом **Конгрессе** занял республиканец Джон Бенер.

Information Extraction Mechanism.

Место спикера **Палаты представителей** в новом **Конгрессе** занял республиканец **Джон Бенер**.

Information Extraction Mechanism.

Место спикера **Палаты представителей** в новом **Конгрессе** занял республиканец **Джон Бенер**.

Occupation

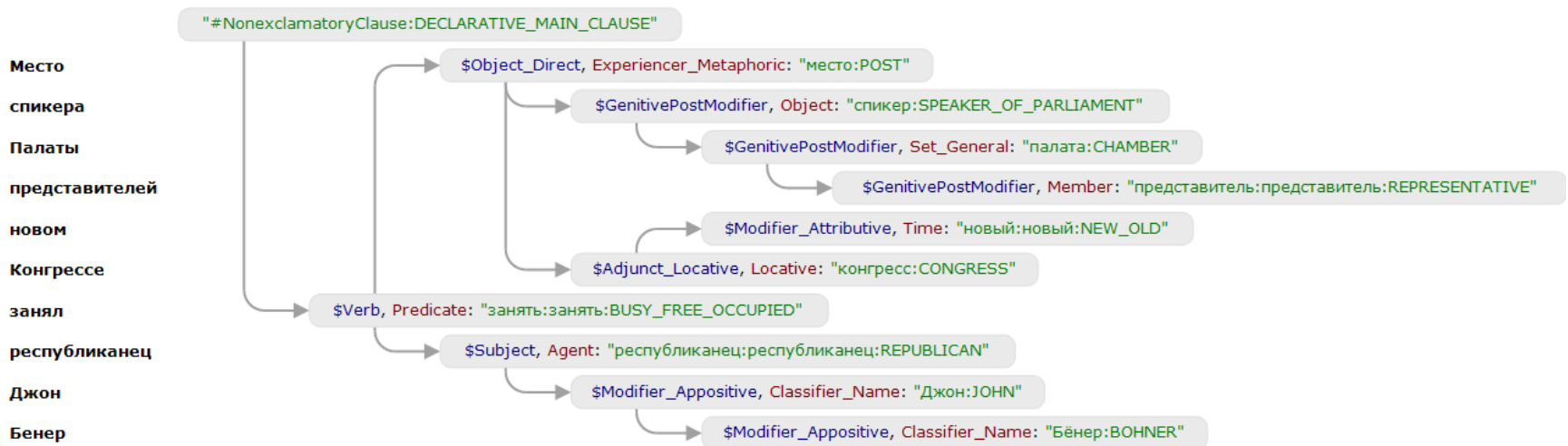
Who: Джон Бенер

Where: Палата представителей

Position: спикер

ABBYY InfoExtractor. Syntactic-Semantic trees

Место спикера Палаты представителей в новом Конгрессе занял республиканец Джон Бенер.



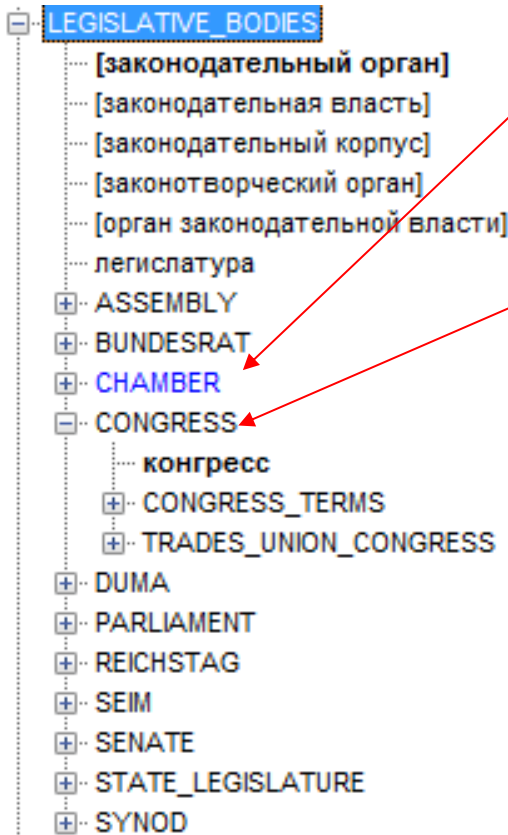
Information Extraction Mechanism.

Interpretation Rules

- Left side
 - a pattern defining a fragment of a syntactic-semantic tree
 - object conditions
 - construction for coreference resolution
- Right side
 - statements that create an object
 - statements that fill object`s attributes
 - statements that create annotations
 - and so on

Information Extraction Mechanism. Interpretation Rules

Место спикера **Палаты представителей** в новом **Конгрессе** занял республиканец Джон Бенер.



Information Extraction Mechanism. Interpretation Rules

Место спикера **Палаты представителей** в новом **Конгрессе** занял республиканец Джон Бенер.

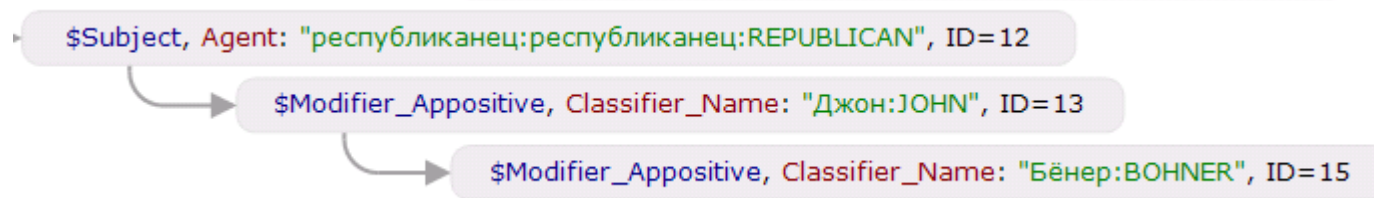
- LEGISLATIVE_BODIES
 - [законодательный орган]
 - [законодательная власть]
 - [законодательный корпус]
 - [законотворческий орган]
 - [орган законодательной власти]
 - легислатура
 - + ASSEMBLY
 - + BUNDESRAT
 - + CHAMBER
 - CONGRESS
 - конгресс
 - + CONGRESS_TERMS
 - + TRADES_UNION_CONGRESS
 - + DUMA
 - + PARLIAMENT
 - + REICHSTAG
 - + SEIM
 - + SENATE
 - + STATE_LEGISLATURE
 - + SYNOD

```

"LEGISLATIVE_BODIES"
=>
Organization Org (this),
annotation(Org, this.core),
Org.pre_identifier == Norm(this.core);
  
```

Information Extraction Mechanism. Interpretation Rules

Место спикера Палаты представителей в новом Конгрессе занял республиканец Джон Бенер.

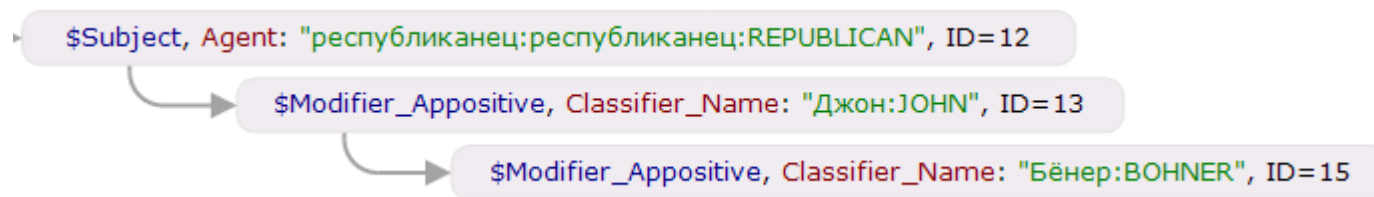


```

//Илья Петров пришел
this "PERSON_BY_FIRSTNAME"
[
  surname "PERSON_BY_LASTNAME"
]
=>
Person P(this),
anchor(P, surname, Coreferential),
annotation(P, this.core, surname.core),
P.firstname == Norm(this.core),
P.surname == Norm(surname.core);
  
```

Information Extraction Mechanism. Interpretation Rules

Место спикера Палаты представителей в новом Конгрессе занял республиканец Джон Бенер.



```

//студент Вася
head "HUMAN"
[
  (Classifier_Name|ClassifiedEntity|Specification): this <*>BasicEntity:Person*>
]
=>
anchor(this.o,head, Coreferential);
  
```

Information Extraction Mechanism. Interpretation Rules

Место спикера Палаты представителей в новом Конгрессе занял республиканец Джон Бенер.

```
v "BUSY_FREE_OCCUPIED"|"TO_OBTAIN_OFFICIAL_POSITION"
[
  occ "SOCIAL_POSITION"|"TITLE"
  [
    this "MAN_AS_PROFESSIONAL_WORKER"
  ]
]
[
  Agent !Per <%Person%>
]
=>
Occupation Occ(this, NoDistribution),
anchor (Occ, v, NoDistribution),
Occ.position == Norm(this.core),
Occ.employee == Per.o;
```

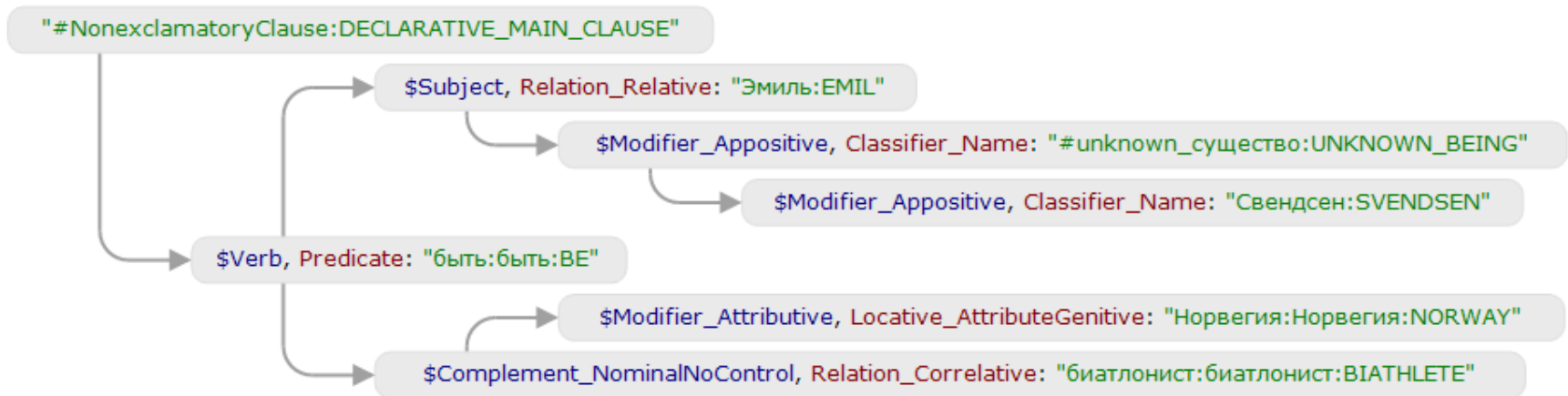
```
this <%Occupation%>
[
  (Set_General | Locative): !Org <%Organization%>
]
=>
this.o.employer == Org.o;
```

Information Extraction Mechanism. Interpretation Rules. Anaphoric Rules

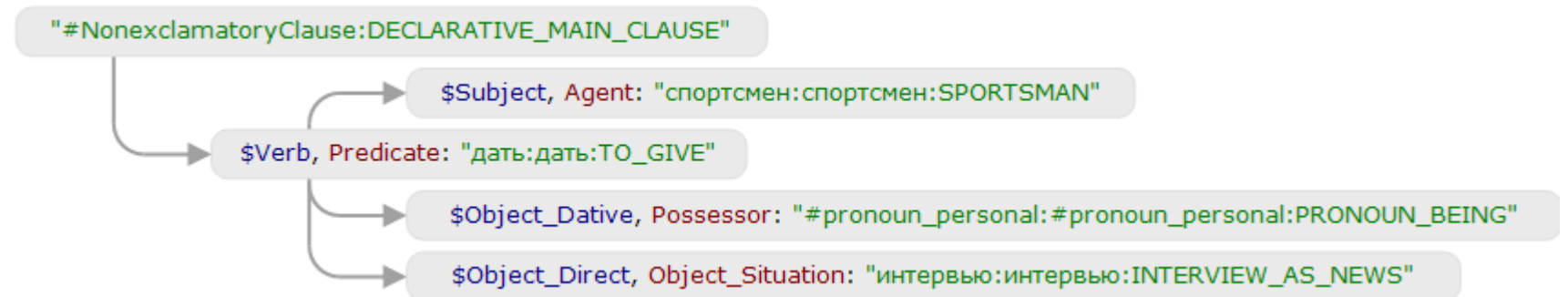
Эмиль Хегле Свендсен - норвежский биатлонист. Спортсмен дал нам интервью.

Information Extraction Mechanism. Interpretation Rules. Anaphoric Rules

Эмиль Хегле Свендсен - норвежский биатлонист.

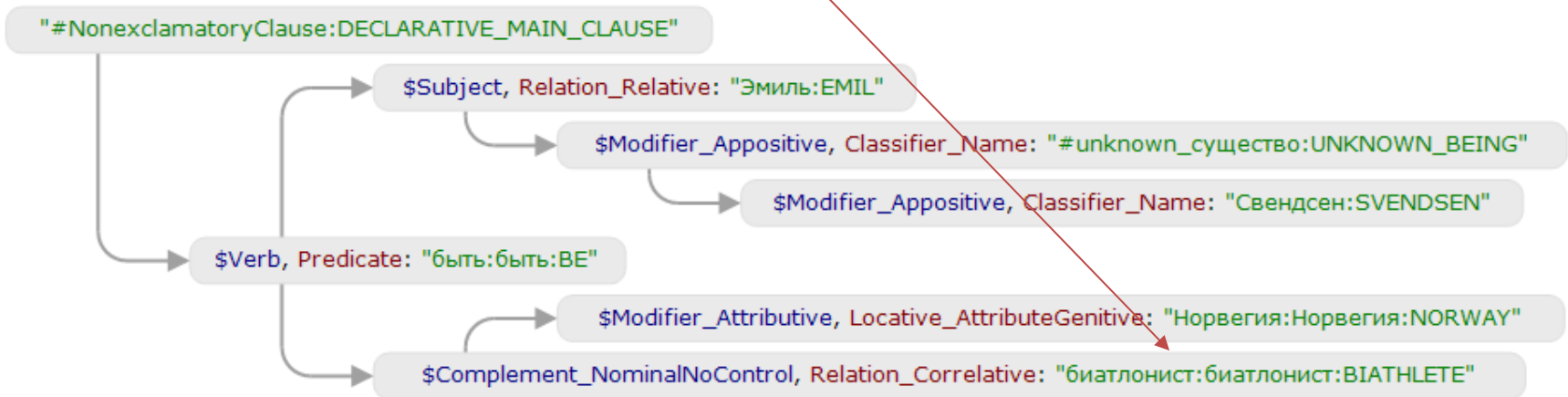


Спортсмен дал нам интервью

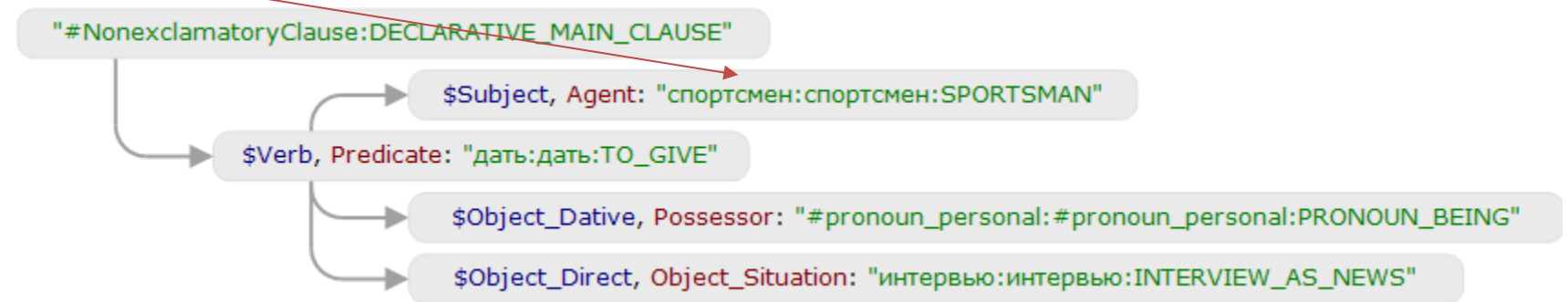


Information Extraction Mechanism. Interpretation Rules. Anaphoric Rules

Эмиль Хегле Свендсен - норвежский биатлонист.



Спортсмен дал нам интервью



Information Extraction Mechanism. Interpretation Rules. Anaphoric Rules

Эмиль Хегле Свендсен - норвежский биатлонист. Спортсмен дал нам интервью.

```

this "HUMAN" <% Person, unnamed %>
<$
  distance: 3, template:
  !person "HUMAN" "this.SemanticClass" <% Person %>
$>
=>
TheSame(this.o, person.o);

```

Information Extraction Mechanism. Identification Rules

- Left side
 - patterns describing information objects (fragment of a graph)
 - conditions
- Right side
 - statement that two objects must be merged

Information Extraction Mechanism.

Identification rules

An example of a rule merging two “Purchase and Sale” facts by “customer” and “property” attributes

```
<% PurchaseAndSale, customer ~= null, property ~= null, seller == null %>
<% PurchaseAndSale, customer ~= null, property ~= null %>

intersects( customer ),
intersects( property )
```

Results. Track 1. (Nickname – violet)

With Loc/LocOrg division

Type	Precision	Recall	F1
per	0.9450	0.9155	0.9300
loc	0.5168	0.8596	0.6455
org	0.8175	0.7564	0.7858
locorg	0.8864	0.3122	0.4618
overall	0.7880	0.7495	0.7682

Without Loc/LocOrg division

Type	Precision	Recall	F1
per	0.9450	0.9155	0.9300
loc	0.9261	0.8698	0.8971
org	0.8175	0.7564	0.7858
overall	0.8931	0.8427	0.8672

Results. Track 2. (Nickname – violet)

Type	Precision	Recall	F1
per	0.8768	0.8572	0.8669
loc	0.8413	0.8135	0.8272
org	0.6757	0.6869	0.6812
OVERALL	0.7860	0.7778	0.7819

Results. Track 3. (Nickname – violet)

Type	Precision	Recall	F1
ownership	0.5379	0.1709	0.2594
occupation	0.8013	0.5646	0.6624
meeting	0.8690	0.1352	0.2340
deal	0.6777	0.1932	0.3007
overall	0.7494	0.3840	0.5078

Thank you!
Questions?