# Coreference Resolution in Russian: State-of-the-Art Approaches Application and Evolvement

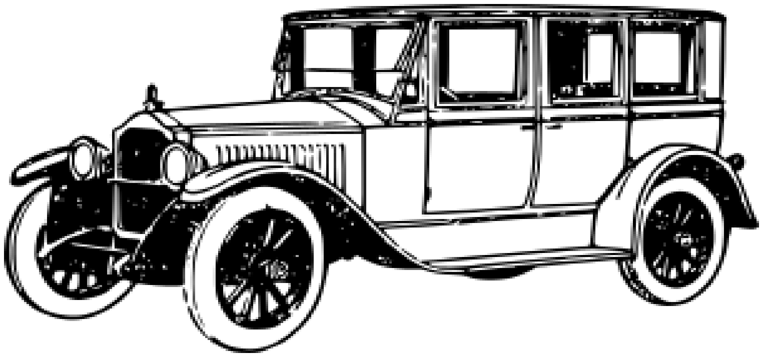
**Speaker: Andrey Sysoev**
Ivan Andrianov
Alexandra Khadzhiiskaia

Moscow, 2017

# Let's start with an example

If a **bulb** in your **car** burned out – change **it**.

Если в **машине** перегорела **лампа** – замените **её**.

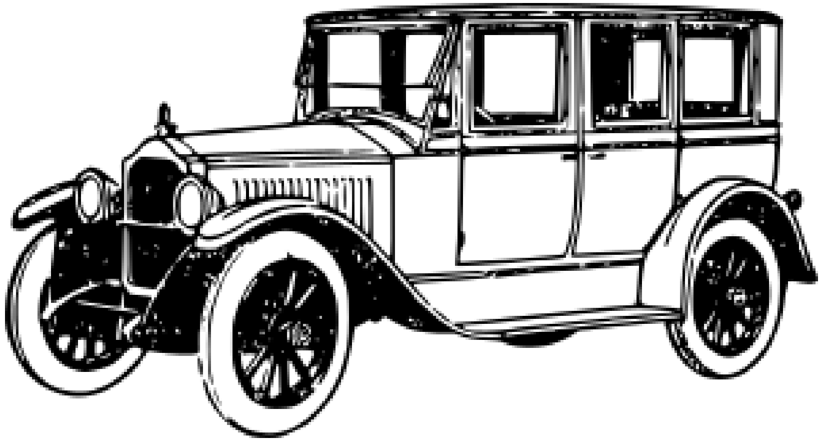


*What should I actually change?*

# Let's start with an example

If a **bulb** in your **car** burned out – change **it**.

Если в **машине** перегорела **лампа** – замените **её**.

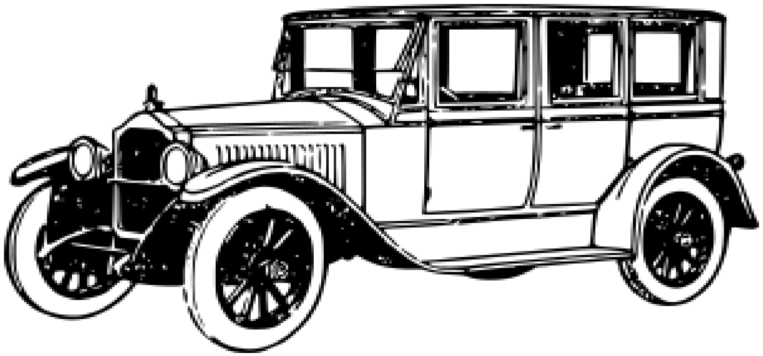


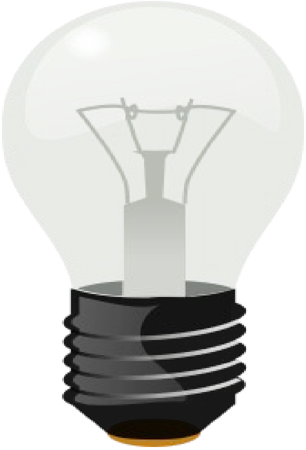
*What should I actually change? A car?*

# Let's start with an example

If a **bulb** in your **car** burned out – change **it**.

Если в **машине** перегорела **лампа** – замените **её**.





*What should I actually change? A car? Or just a bulb?*

# Coreference resolution. Where to use?

- Relation extraction
- Question-answering systems
- Sentiment analysis

# Coreference resolution in practice

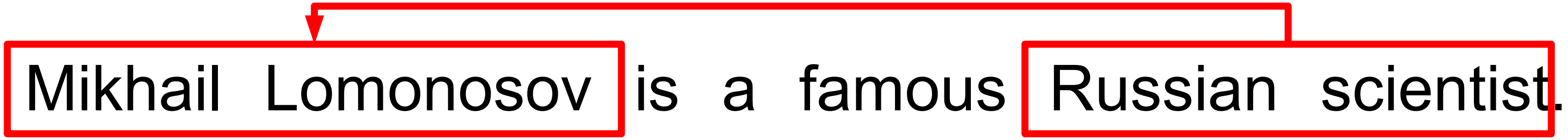
Mikhail Lomonosov is a famous Russian scientist.

One of his discoveries is the atmosphere of Venus.

Михаил Васильевич Ломоносов – выдающийся русский ученый.

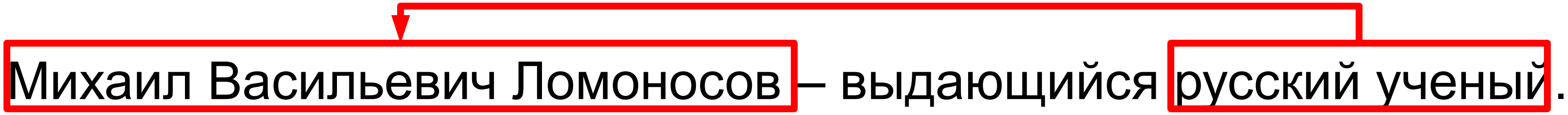
Одно из его открытий – атмосфера Венеры.

# Coreference resolution in practice



Mikhail Lomonosov is a famous Russian scientist.

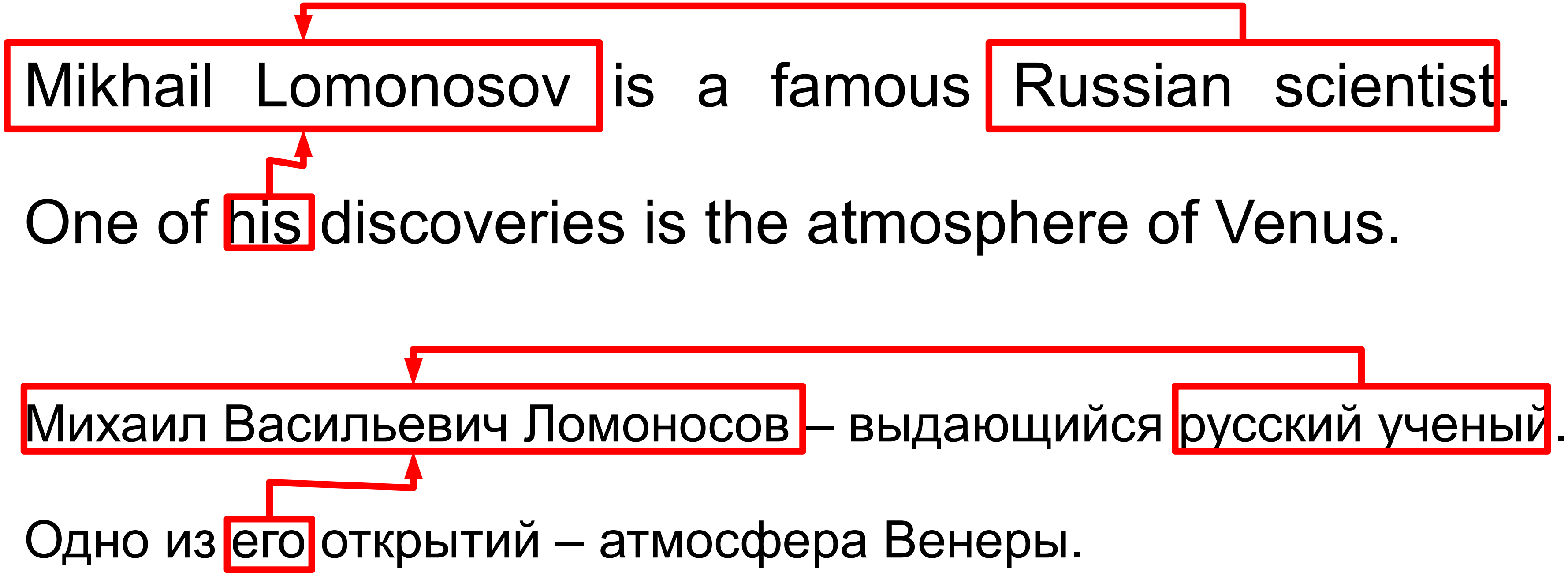
One of his discoveries is the atmosphere of Venus.



Михаил Васильевич Ломоносов – выдающийся русский ученый.

Одно из его открытий – атмосфера Венеры.

# Coreference resolution in practice



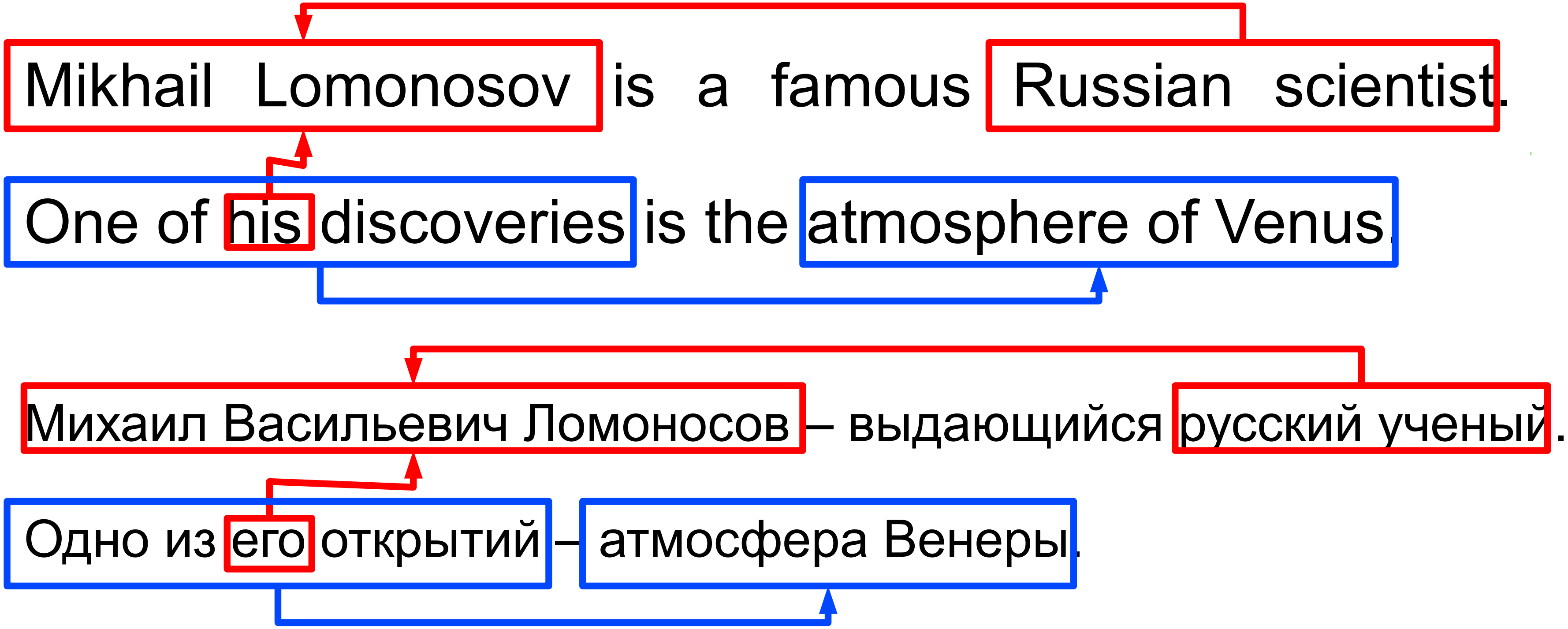
Mikhail Lomonosov is a famous Russian scientist.

One of his discoveries is the atmosphere of Venus.

Михаил Васильевич Ломоносов – выдающийся русский ученый.

Одно из его открытий – атмосфера Венеры.

# Coreference resolution in practice



Mikhail Lomonosov is a famous Russian scientist.

One of his discoveries is the atmosphere of Venus.

Михаил Васильевич Ломоносов – выдающийся русский ученый.

Одно из его открытий – атмосфера Венеры.

# Coreference resolution in practice

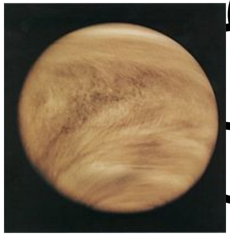
Mikhail Lomonosov is a famous Russian scientist.

One of his discoveries is the atmosphere of Venus.

Михаил Васильевич Ломоносов – выдающийся русский ученый.

Одно из его открытий – атмосфера Венеры.

Mikhail Lomonosov
Russian scientist
his



One of his discoveries
atmosphere of Venus

# Antecedent and anaphor

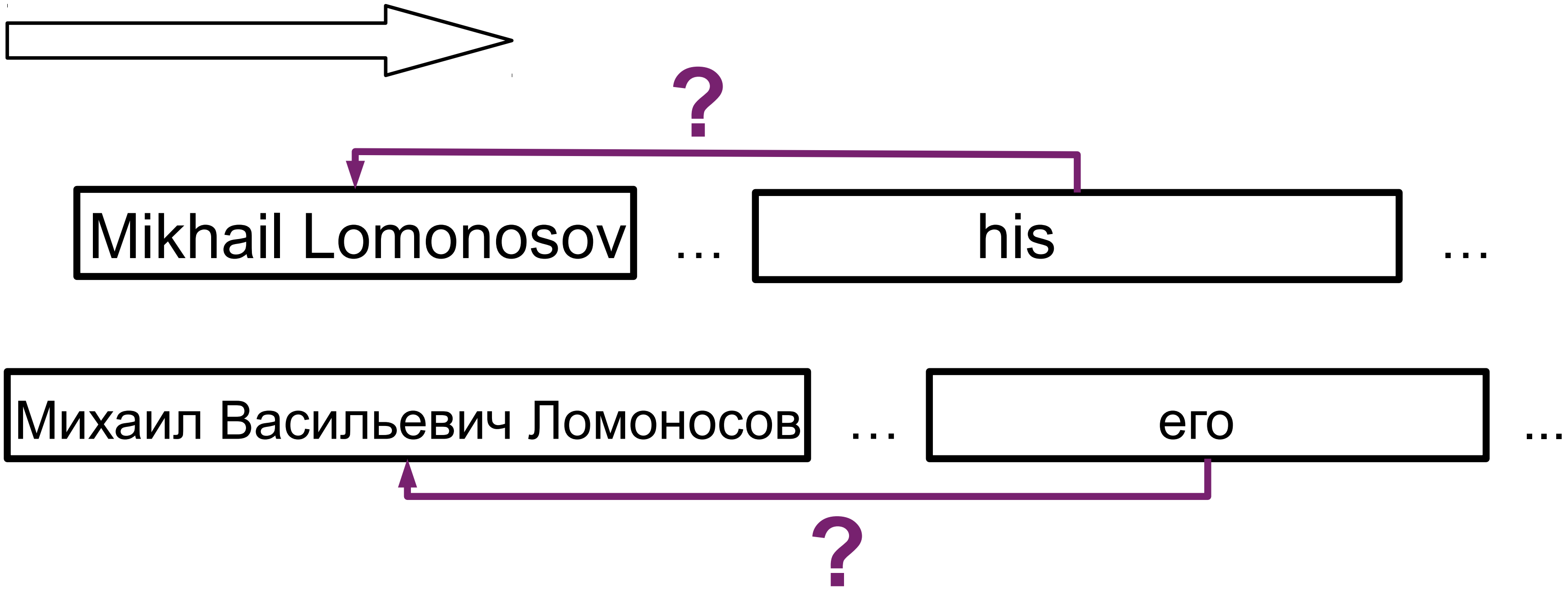
Antecedent - the mention, which already has some meaning (within the text).

Anaphor - the mention, which borrows its meaning from corresponding antecedent.

# Antecedent and anaphor

Antecedent - the mention, which already has some meaning (within the text).

Anaphor - the mention, which borrows its meaning from corresponding antecedent.



?

Mikhail Lomonosov … his …

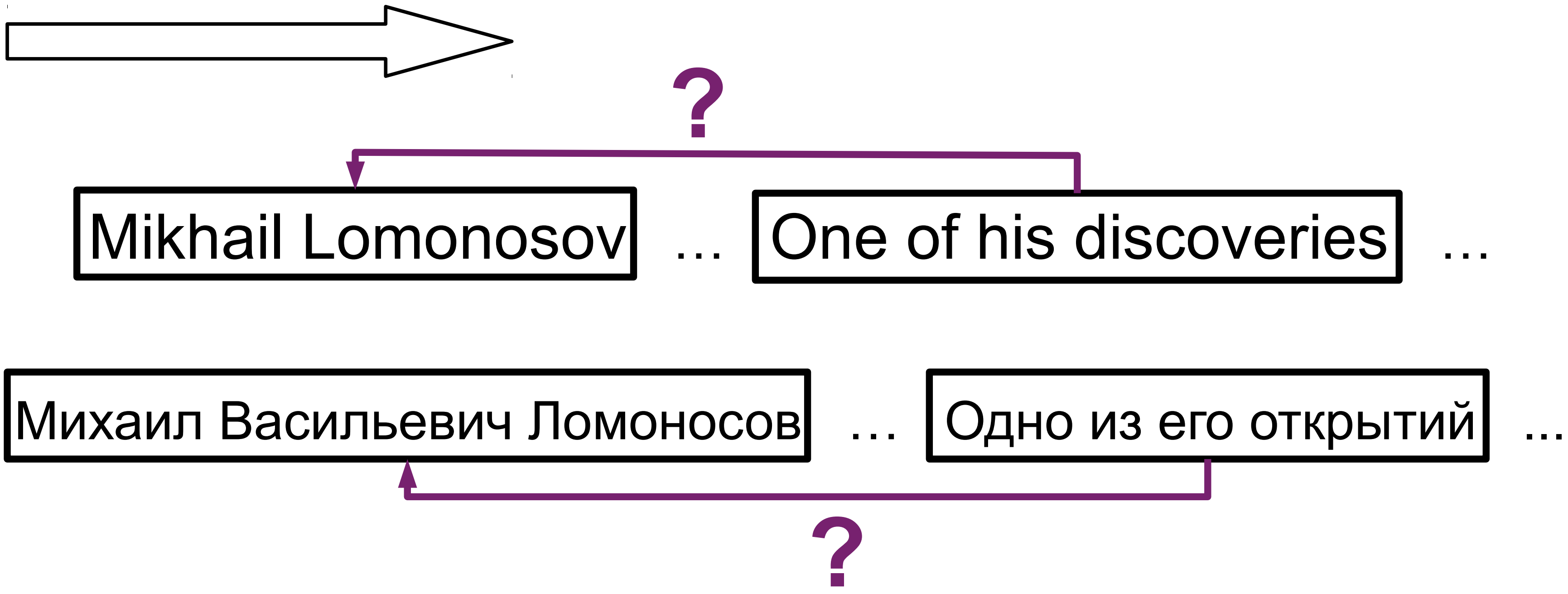
Михаил Васильевич Ломоносов … его ...

?

# Antecedent and anaphor

Antecedent - the mention, which already has some meaning (within the text).

Anaphor - the mention, which borrows its meaning from corresponding antecedent.



?

Mikhail Lomonosov … One of his discoveries …

Михаил Васильевич Ломоносов … Одно из его открытий ...

?

# Evaluation.
# Test corpus and metrics

Corpus:

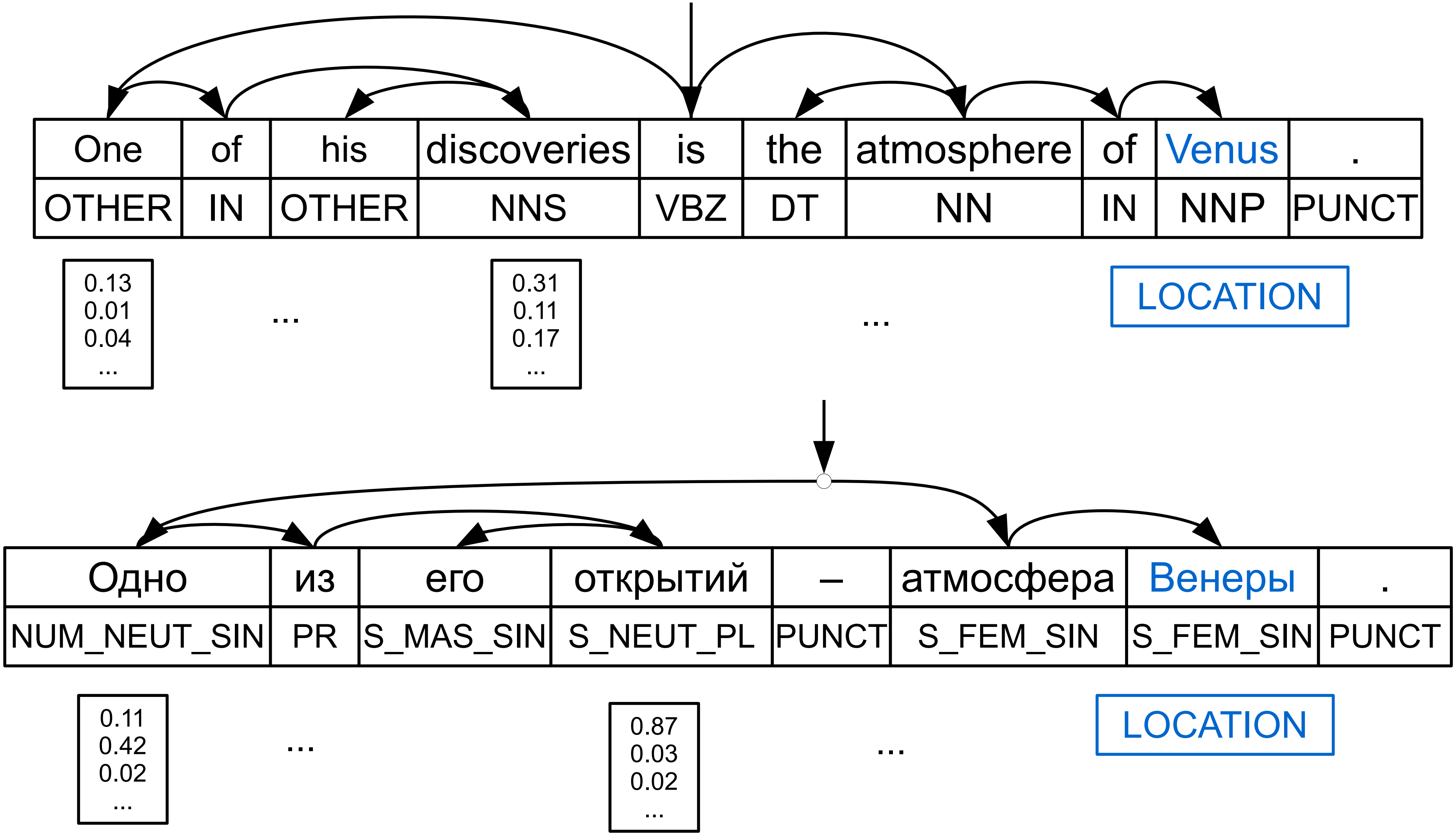
- RuCor from RuEval-2014
- 181 (179 after fixing conflicting markup) documents

Quality measures:

- Precision / Recall / F1
- MUC / B3 / $\text{CEAF}_{\text{entity}}$ / $\text{CEAF}_{\text{mention}}$

10-fold crossvalidation

# Preprocessing



| One | of | his | discoveries | is | the | atmosphere | of | Venus | . |
|---|---|---|---|---|---|---|---|---|---|
| OTHER | IN | OTHER | NNS | VBZ | DT | NN | IN | NNP | PUNCT |

0.13<br>0.01<br>0.04<br>...

...

0.31
0.11
0.17
...

...

LOCATION

| Одно | из | его | открытий | – | атмосфера | Венеры | . |
|---|---|---|---|---|---|---|---|
| NUM_NEUT_SIN | PR | S_MAS_SIN | S_NEUT_PL | PUNCT | S_FEM_SIN | S_FEM_SIN | PUNCT |

0.11
0.42
0.02
...

...

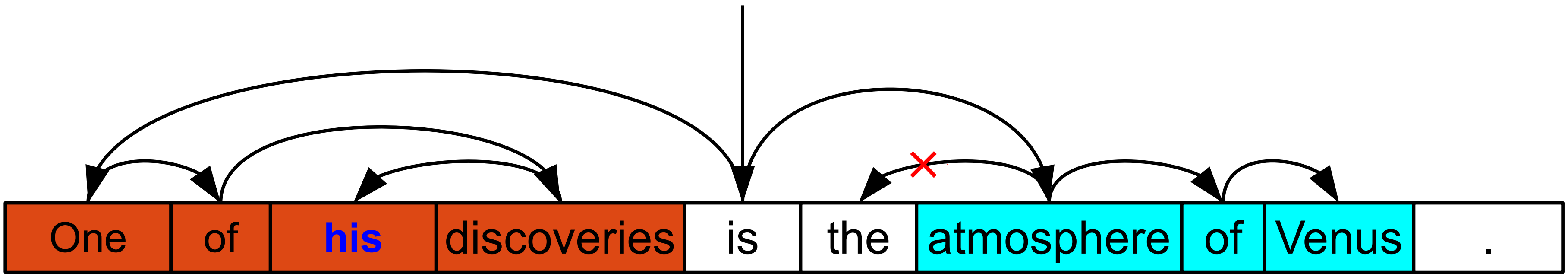
0.87
0.03
0.02
...

...

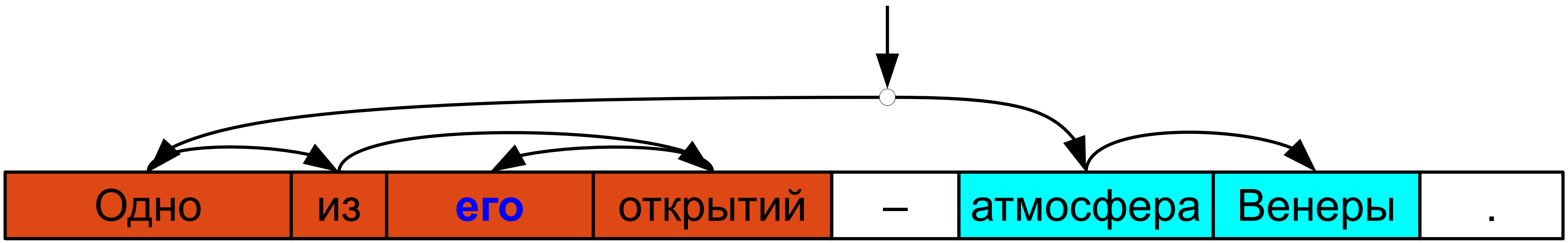
LOCATION

*Texterra: https://api.ispras.ru/demo/texterra*

# Mention detection.
# What is mention?

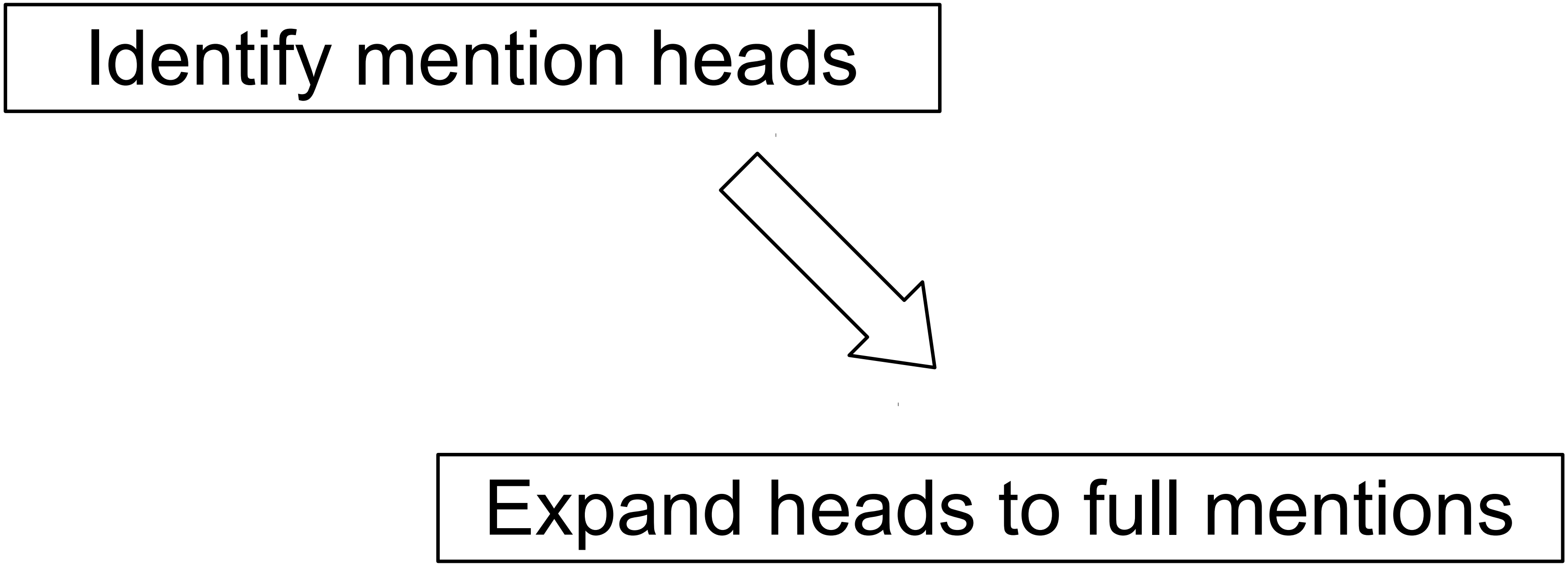


One of **his** discoveries is the atmosphere of Venus .



Одно из **его** открытий – атмосфера Венеры .

# Mention detection

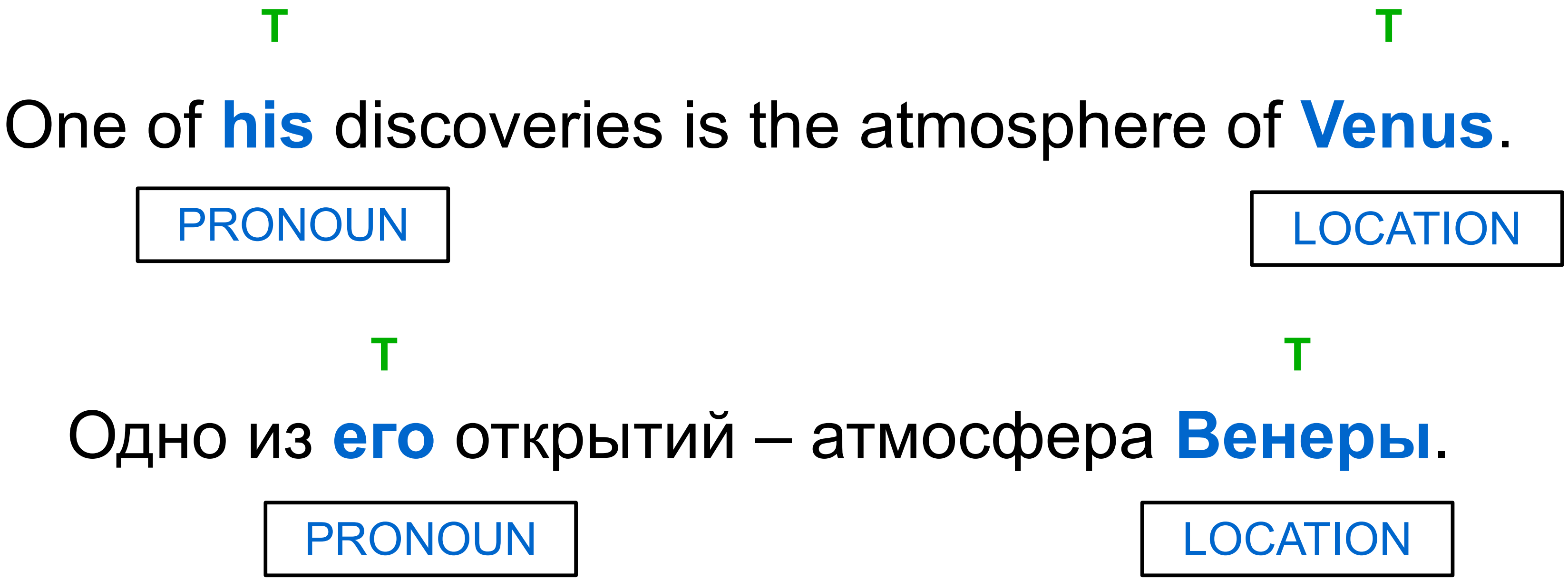


Identify mention heads

Expand heads to full mentions

*Peng H., Chang K.-W., Roth D. (2015)*
*A Joint Framework for Coreference Resolution and Mention Head Detection*

# Head identification



T T

One of **his** discoveries is the atmosphere of **Venus**.

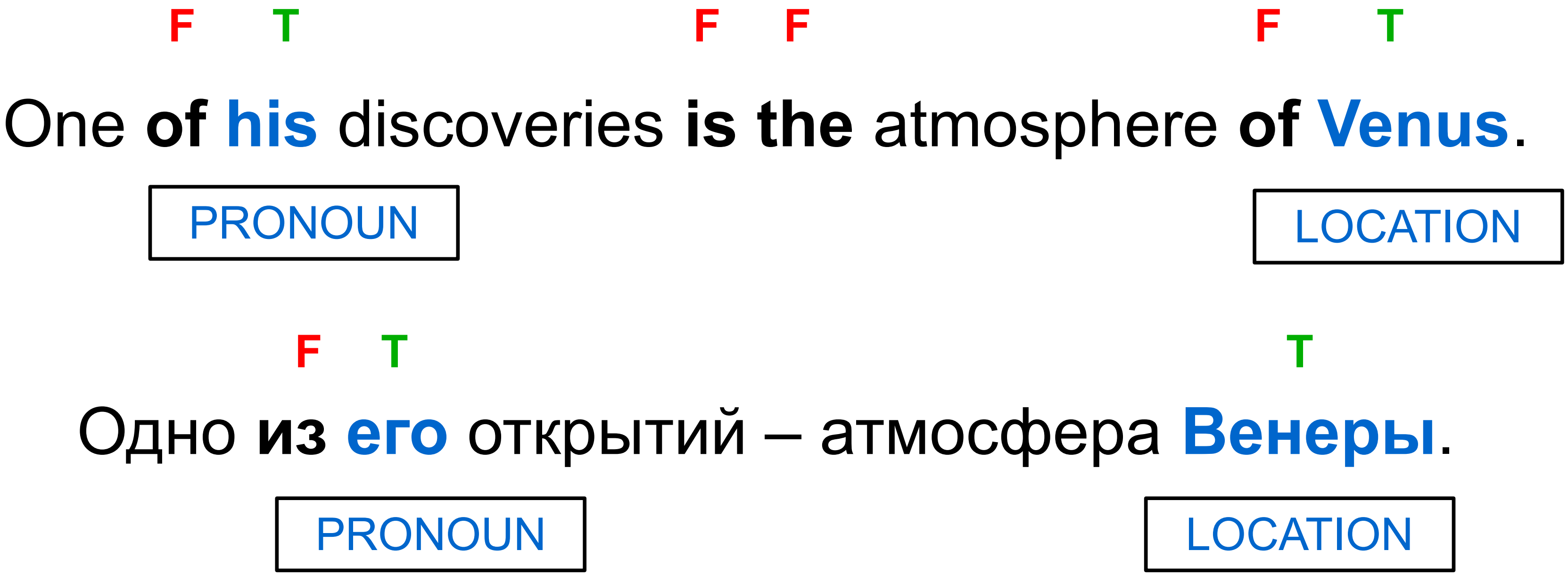
PRONOUN LOCATION

T T

Одно из **его** открытий – атмосфера **Венеры**.

PRONOUN LOCATION

# Head identification



F T F F F T

One **of his** discoveries **is the** atmosphere **of Venus**.

PRONOUN

LOCATION

F T T

Одно **из его** открытий – атмосфера **Венеры**.

PRONOUN

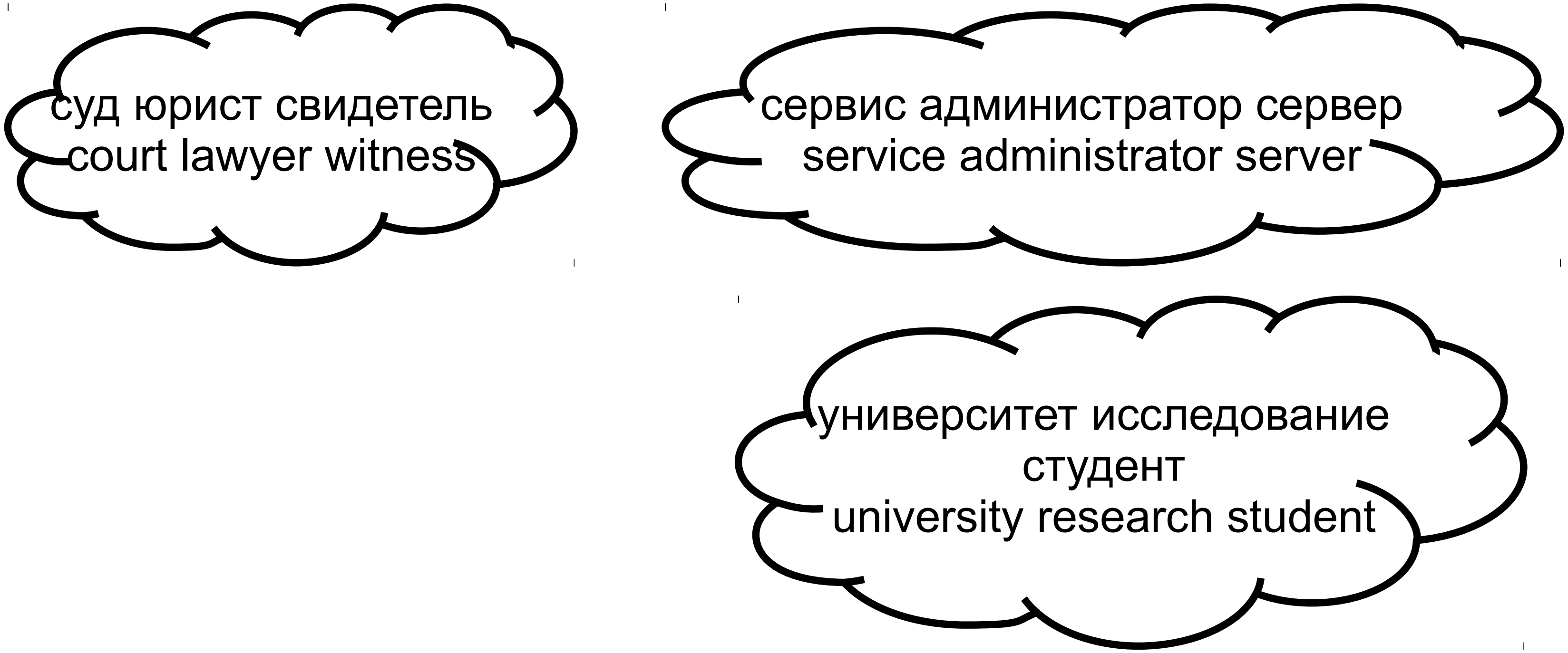
LOCATION

# Head identification. Features

- Internal morphological
  - POS-tag
  - number
  - gender
  - animacy
- Syntax
  - position within sentence
  - relations of a token
- Syntactic context - morphological features for syntactic parent
- Context - basic morphological features for neighbors
- Frequency - TF weighting
- **Semantic**

# Head identification. Semantic features

- Groundtruth heads from training documents are clustered (word2vec, k-means).



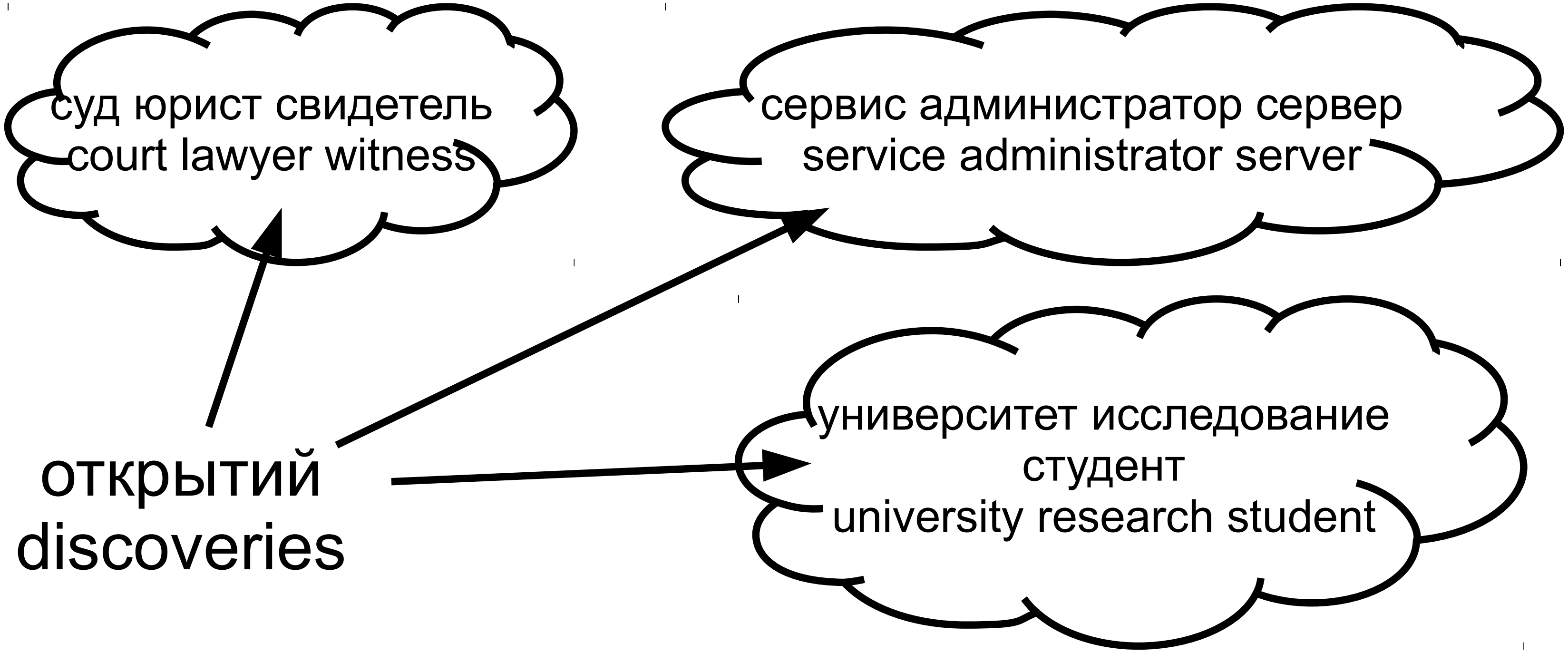
суд юрист свидетель
court lawyer witness

сервис администратор сервер
service administrator server

университет исследование студент
university research student

# Head identification. Semantic features

- Groundtruth heads from training documents are clustered (word2vec, k-means).
- Features - distance and similarity from head candidate to clusters cenroids.



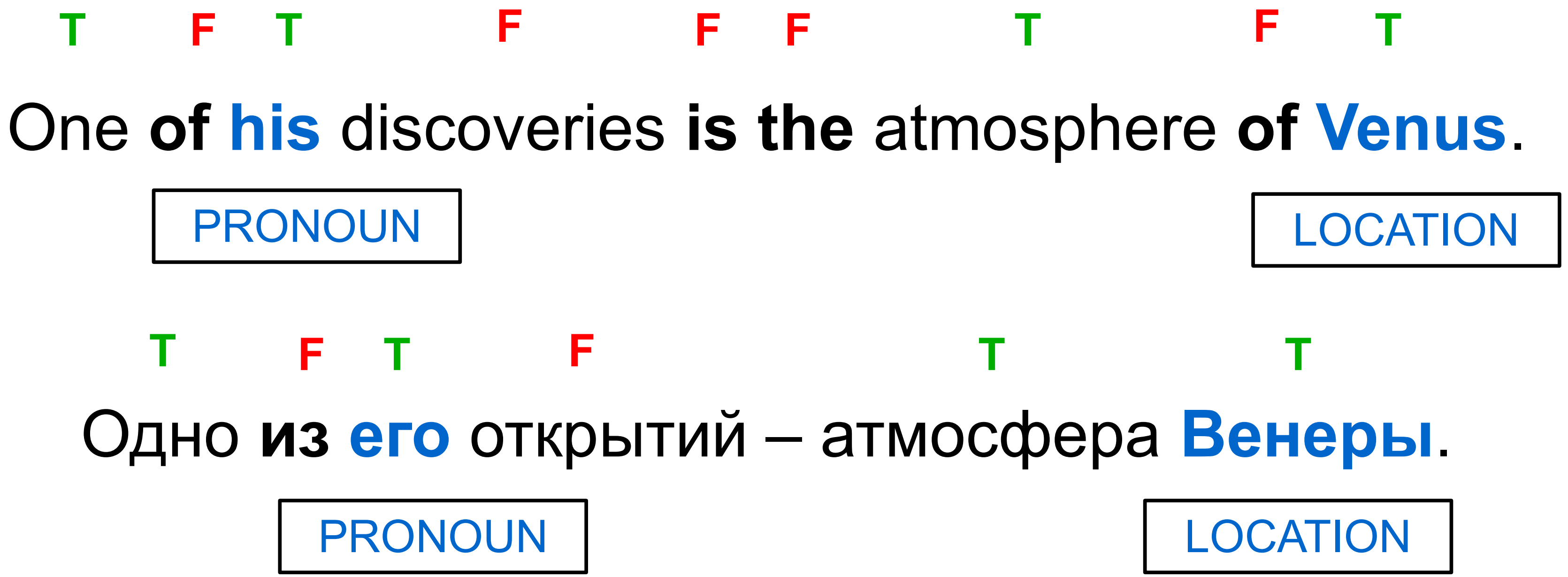
суд юрист свидетель
court lawyer witness

сервис администратор сервер
service administrator server

университет исследование студент
university research student

открытий
discoveries

# Head identification



T F T F F F T F T

One **of his** discoveries **is the** atmosphere **of Venus**.

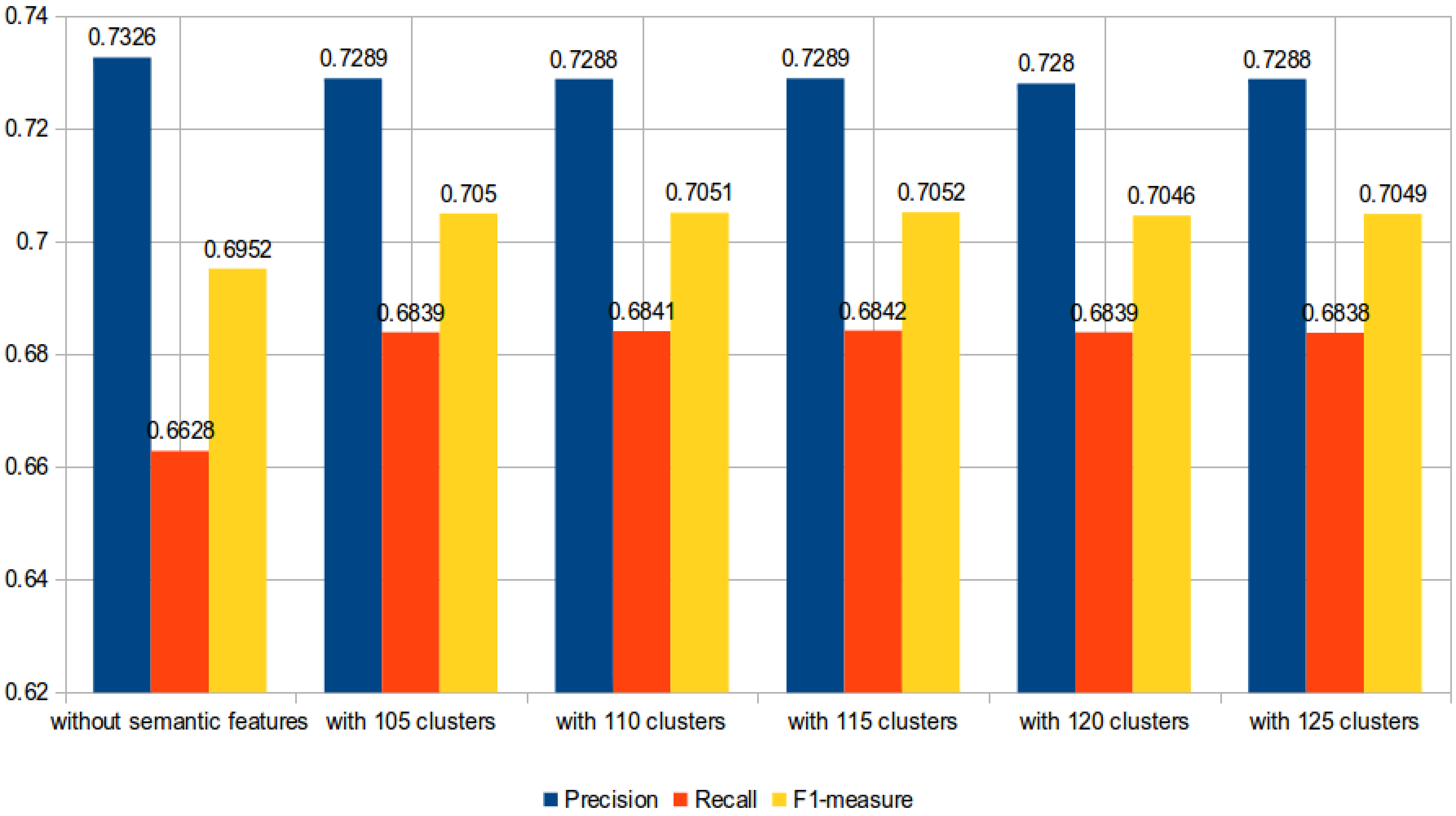
PRONOUN LOCATION

T F T F T T

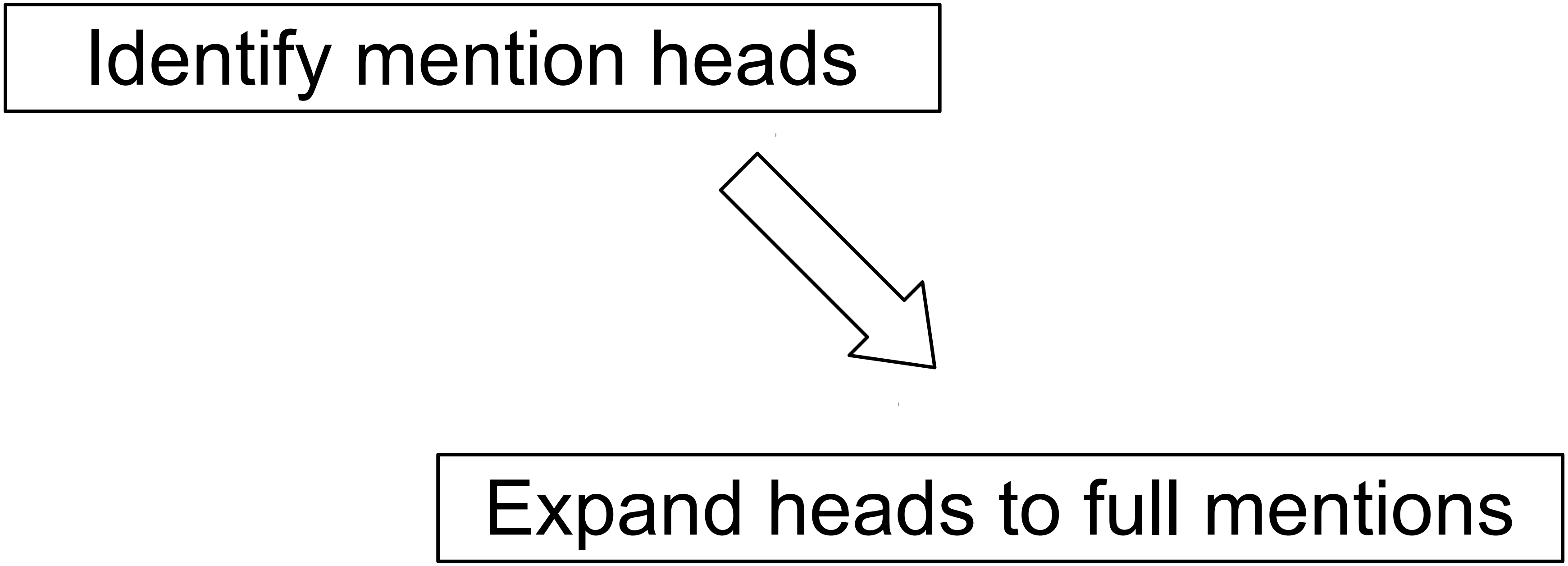
Одно **из его** открытий – атмосфера **Венеры**.

PRONOUN LOCATION

# Head identification. Evaluation



# Mention detection



Identify mention heads

Expand heads to full mentions

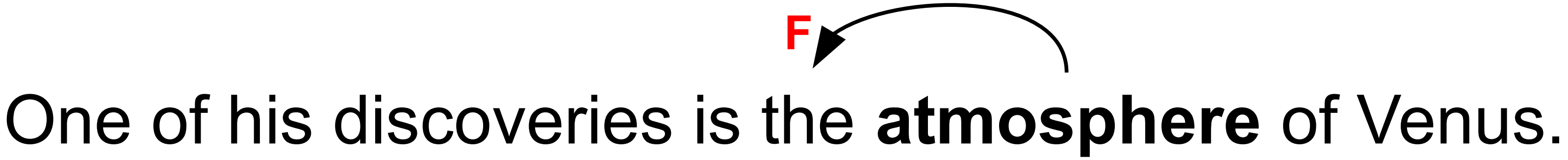
*Peng H., Chang K.-W., Roth D. (2015)*
*A Joint Framework for Coreference Resolution and Mention Head Detection*

# Head expansion

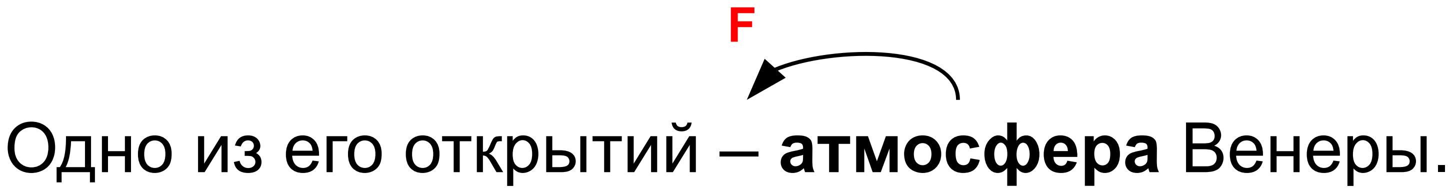
One of his discoveries is the **atmosphere** of Venus.

Одно из его открытий – **атмосфера** Венеры.

# Head expansion

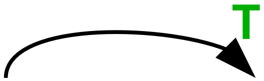


One of his discoveries is the **atmosphere** of Venus.



Одно из его открытий – **атмосфера** Венеры.

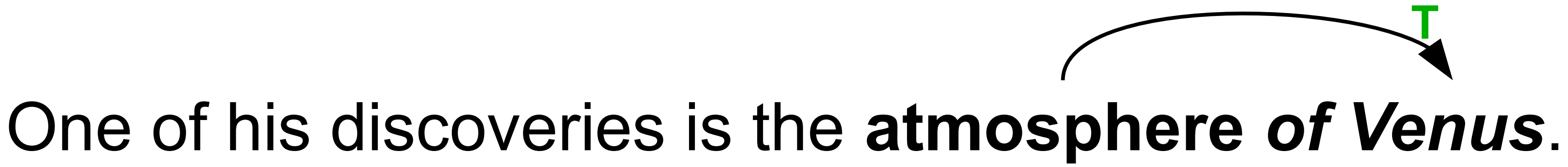
# Head expansion



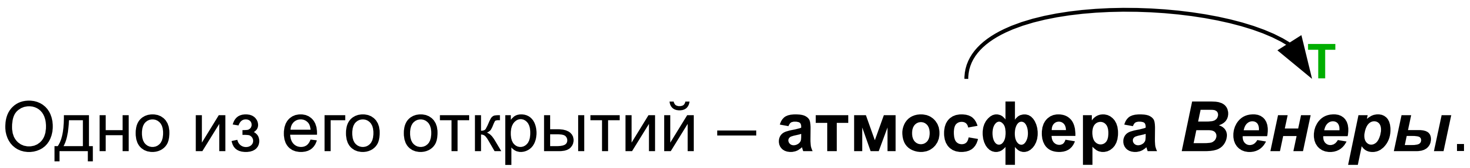
One of his discoveries is the **atmosphere *of*** Venus.

Одно из его открытий – **атмосфера** Венеры.

# Head expansion

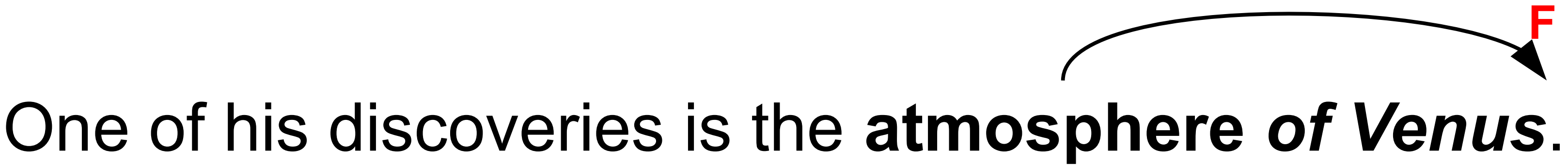


One of his discoveries is the **atmosphere *of Venus***.

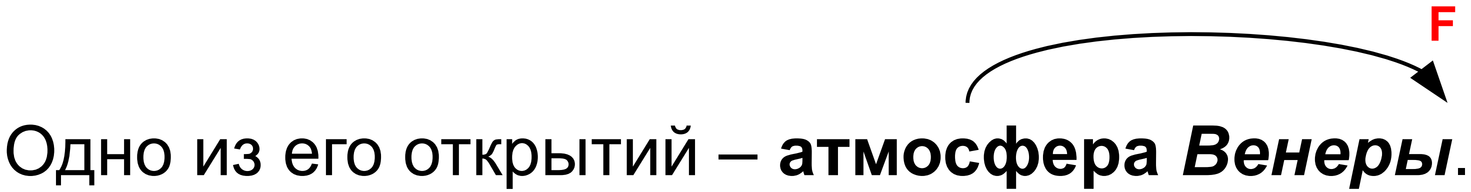


Одно из его открытий – **атмосфера *Венеры***.

# Head expansion



One of his discoveries is the **atmosphere** ***of Venus***.

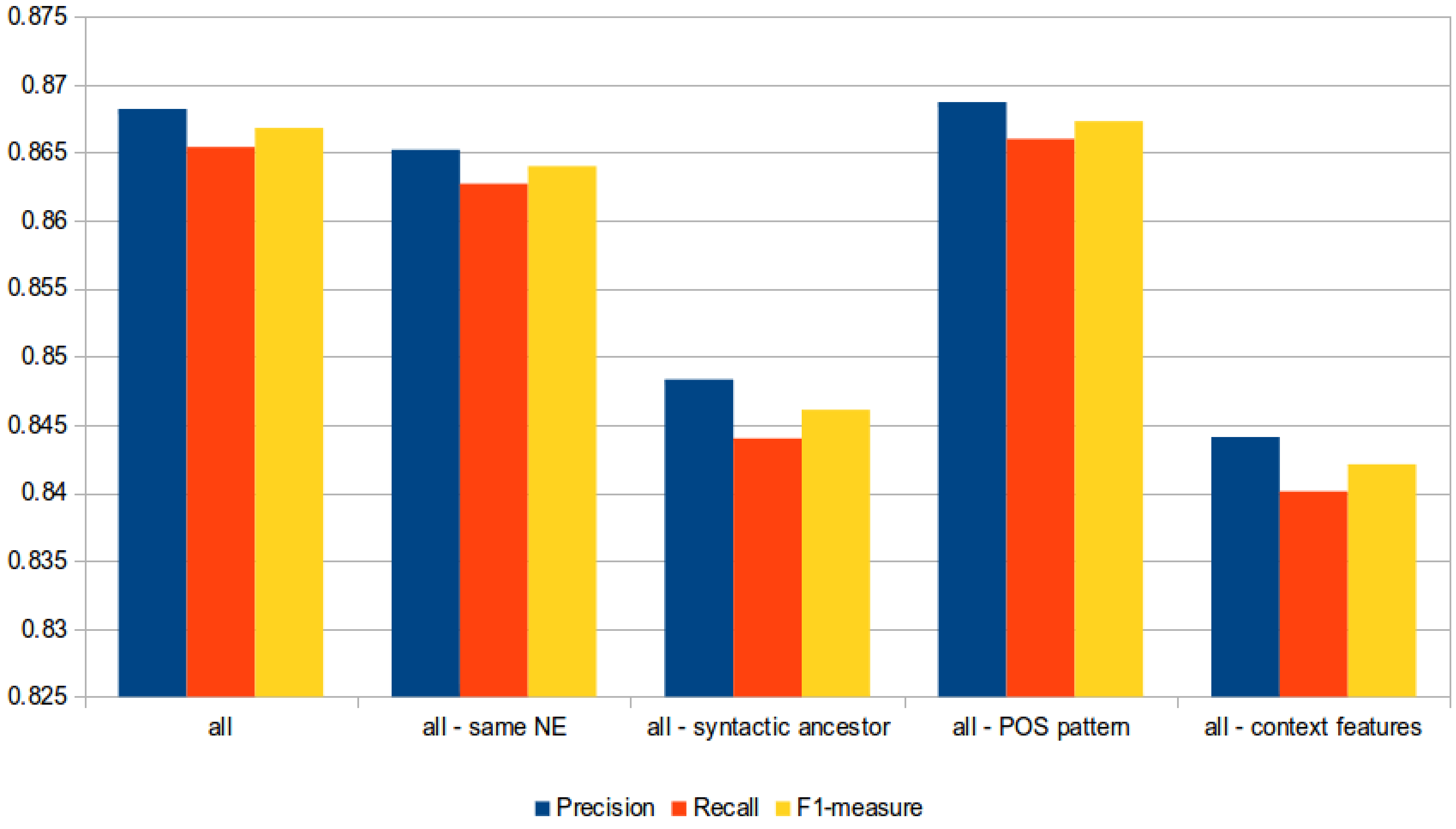


Одно из его открытий – **атмосфера** ***Венеры***.

# Head expansion. Features

- Token-based [head, candidate token, nearest neighbours]
  - word form
  - lemma
  - POS-tag
- Position-based
  - direction from head to candidate
  - distance between head and can-didate
  - head/candidate is the first/last token of the sentence
- Context-based
  - head and candidate are parts of the same named entity
  - head/candidate is a syntactic ancestor of candidate/head
  - POS-tag pattern for words between head and candidate

# Head expansion. Evaluation



# Coreference resolution.
# Easy-First Mention Pair algorithm

*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

# Coreference resolution. Easy-First Mention Pair algorithm

Classified pairs (antecedent-anaphor)

Mikhail Lomonosov - Russian scientist
One of his discoveries - his
One of his discoveries - atmosphere of Venus
Mikhail Lomonosov - his
Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus

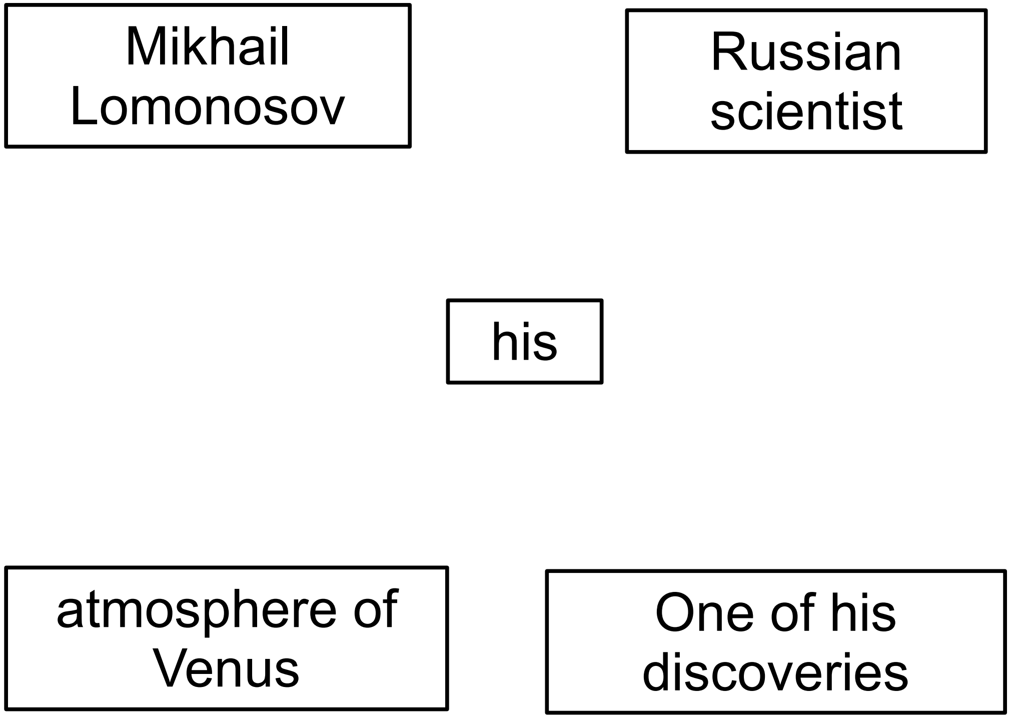
*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

# Coreference resolution.
# Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)

Mikhail Lomonosov - Russian scientist
One of his discoveries - his
One of his discoveries - atmosphere of Venus
Mikhail Lomonosov - his
Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus



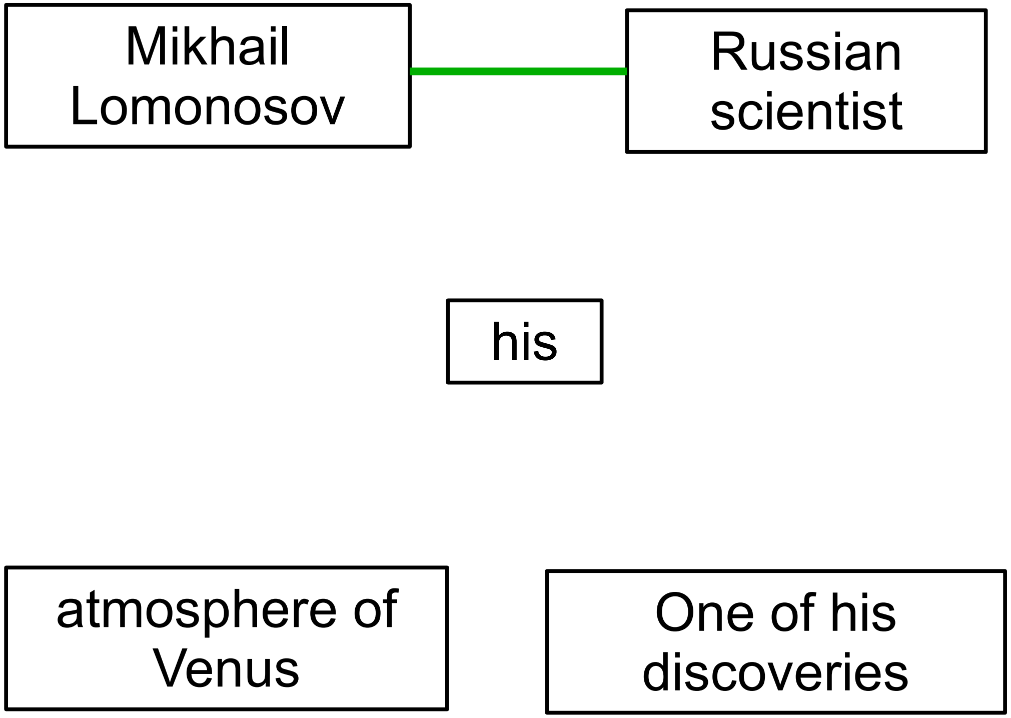
*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

# Coreference resolution.
# Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)

One of his discoveries - his
One of his discoveries - atmosphere of Venus
Mikhail Lomonosov - his
Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus



Mikhail Lomonosov

Russian scientist

his

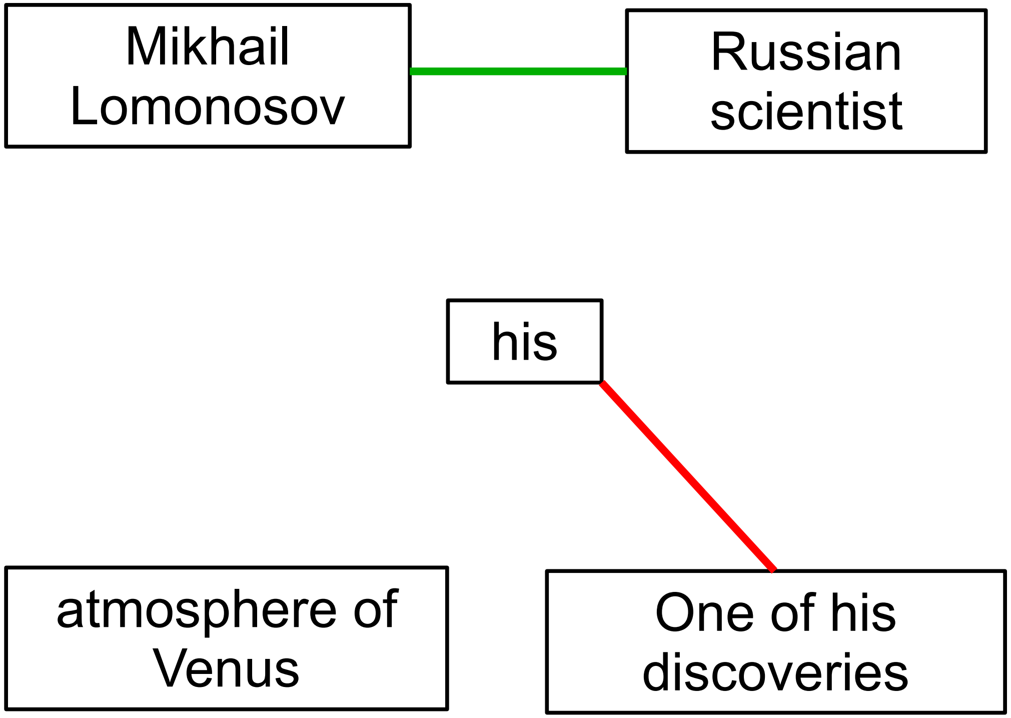
atmosphere of Venus

One of his discoveries

*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

# Coreference resolution.
# Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)



Mikhail Lomonosov

Russian scientist

his

atmosphere of Venus

One of his discoveries

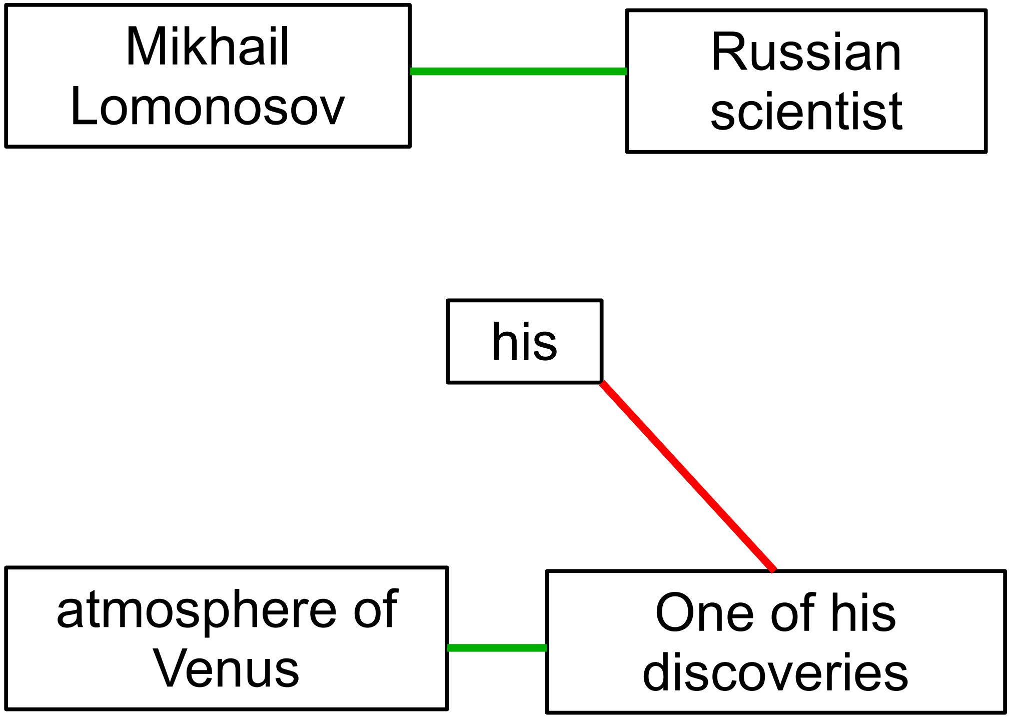
One of his discoveries - atmosphere of Venus
Mikhail Lomonosov - his
Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus

*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

# Coreference resolution.
# Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)



Mikhail Lomonosov - his
Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus

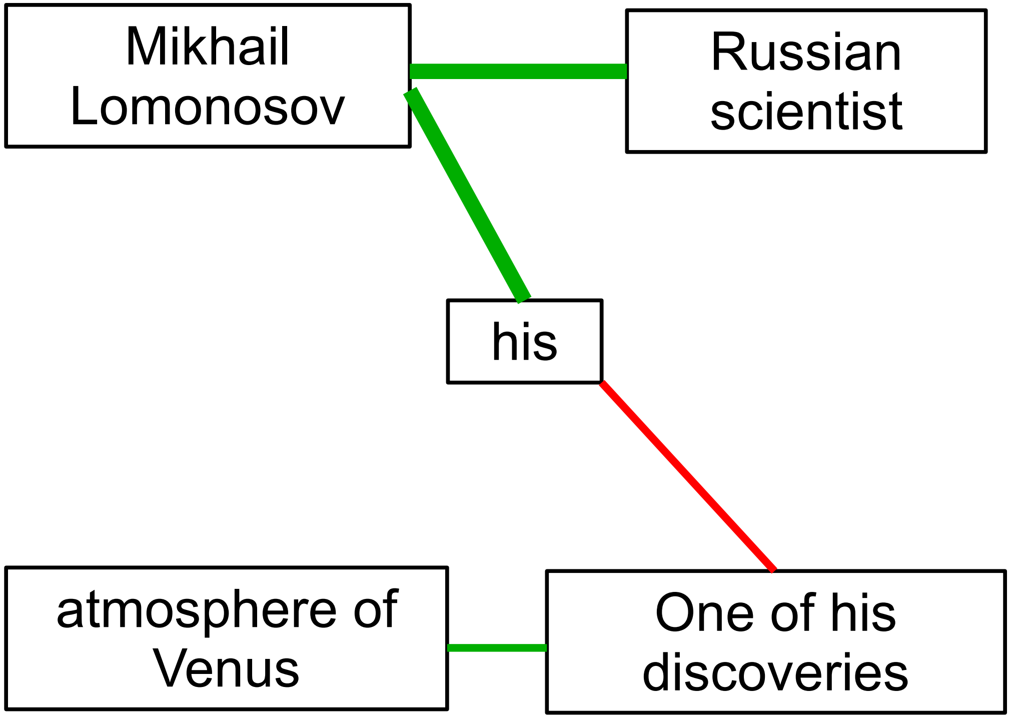
*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

# Coreference resolution. Easy-First Mention Pair algorithm

Classified pairs (antecedent-anaphor)

Russian scientist – his
Mikhail Lomonosov - atmosphere of Venus

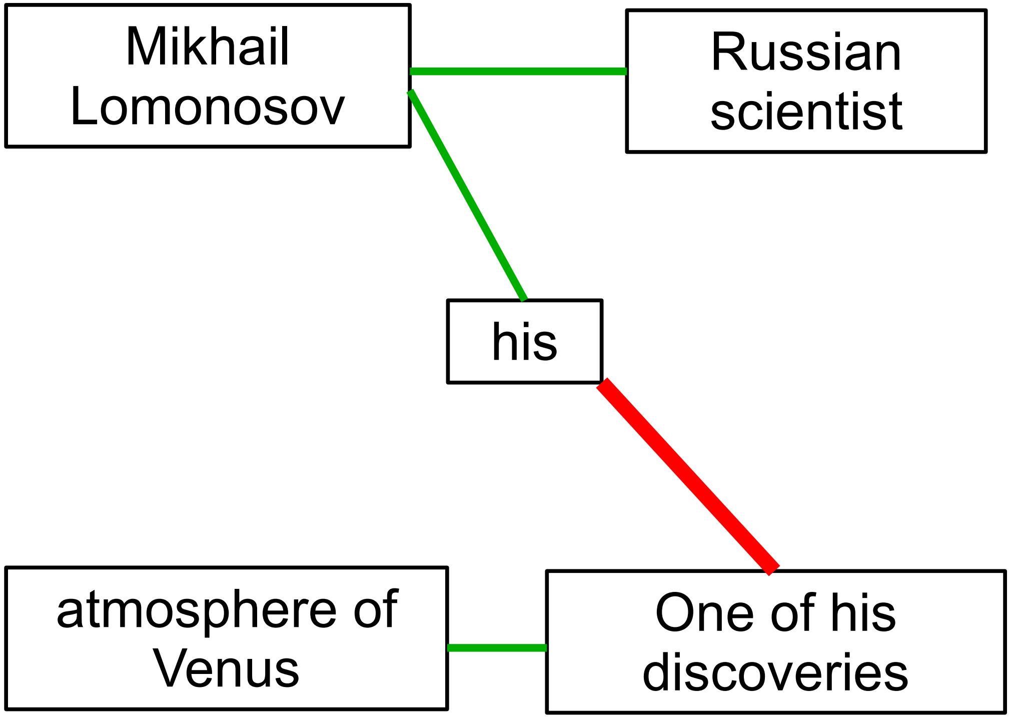
Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus



*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

# Coreference resolution.
# Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)



Mikhail Lomonosov

Russian scientist

his

atmosphere of Venus

One of his discoveries

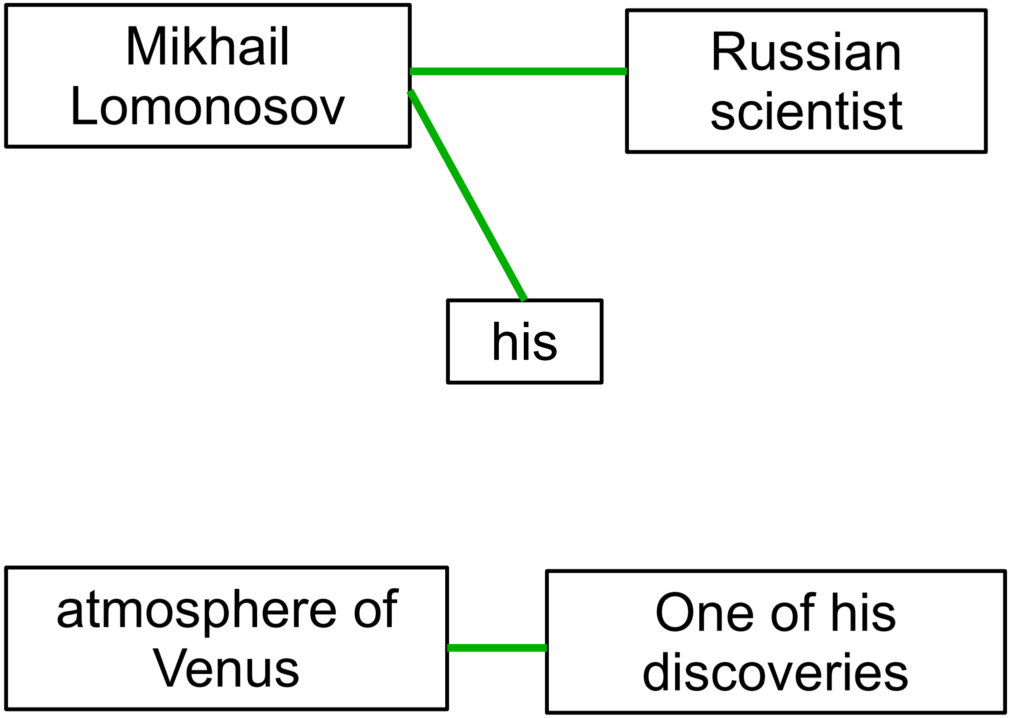
Mikhail Lomonosov - atmosphere of Venus

Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus

*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

# Coreference resolution.
# Easy-First Mention Pair algorithm

Classified pairs
(antecedent-anaphor)



Mikhail Lomonosov - One of his discoveries
Russian scientist - One of his discoveries
Russian scientist - atmosphere of Venus
his - atmosphere of Venus

*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

# Coreference resolution. Mention-pair classifier features

- Basic linguistic: word forms, lemmas, part-of-speech tags, grammemes (gender, number, animacy) for mention head and context words.
- Grammemes agreement: mention heads share the same key grammemes (number, gender, animacy, pro-nominality).
- Positional: distance, place within sentence boundaries.
- Named entity: mention types and their agreement.
- Structural: mention size and interrelation with other mentions of the text.
- Surface form matching: lexicographic similarity and textual representation equality indicators.
- Syntactic: grammar role, sharing same parent node or clause.

# Jaccard Item Set mining

*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

*Segond M., Borgelt C. (2011)*
*Item Set Mining Based on Cover Similarity*

# Jaccard Item Set mining

| | ANT_PER | ANAF_PER | LEX_SIM>0.5 | NUM | ANIM | SENT_ST |
|---|---|---|---|---|---|---|
| *TRUE* | F | F | T | F | F | T |
| *FALSE* | T | T | F | F | T | F |
| *TRUE* | T | F | F | F | T | T |
| *FALSE* | F | F | F | T | F | F |
| *TRUE* | T | F | T | F | T | T |
| *FALSE* | T | T | F | F | T | T |
| *TRUE* | F | T | T | F | F | F |

*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

*Segond M., Borgelt C. (2011)*
*Item Set Mining Based on Cover Similarity*

# Jaccard Item Set mining

| | ANT_PER | ANAF_PER | LEX_SIM>0.5 | NUM | ANIM | SENT_ST |
|---|---|---|---|---|---|---|
| *TRUE* | F | F | T | F | F | T |
| *FALSE* | **T** | **T** | **F** | F | T | F |
| *TRUE* | T | F | F | F | T | T |
| *FALSE* | F | F | F | T | F | F |
| *TRUE* | T | F | T | F | T | T |
| *FALSE* | **T** | **T** | **F** | F | T | T |
| *TRUE* | F | T | T | F | F | F |

*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

*Segond M., Borgelt C. (2011)*
*Item Set Mining Based on Cover Similarity*

# Jaccard Item Set mining

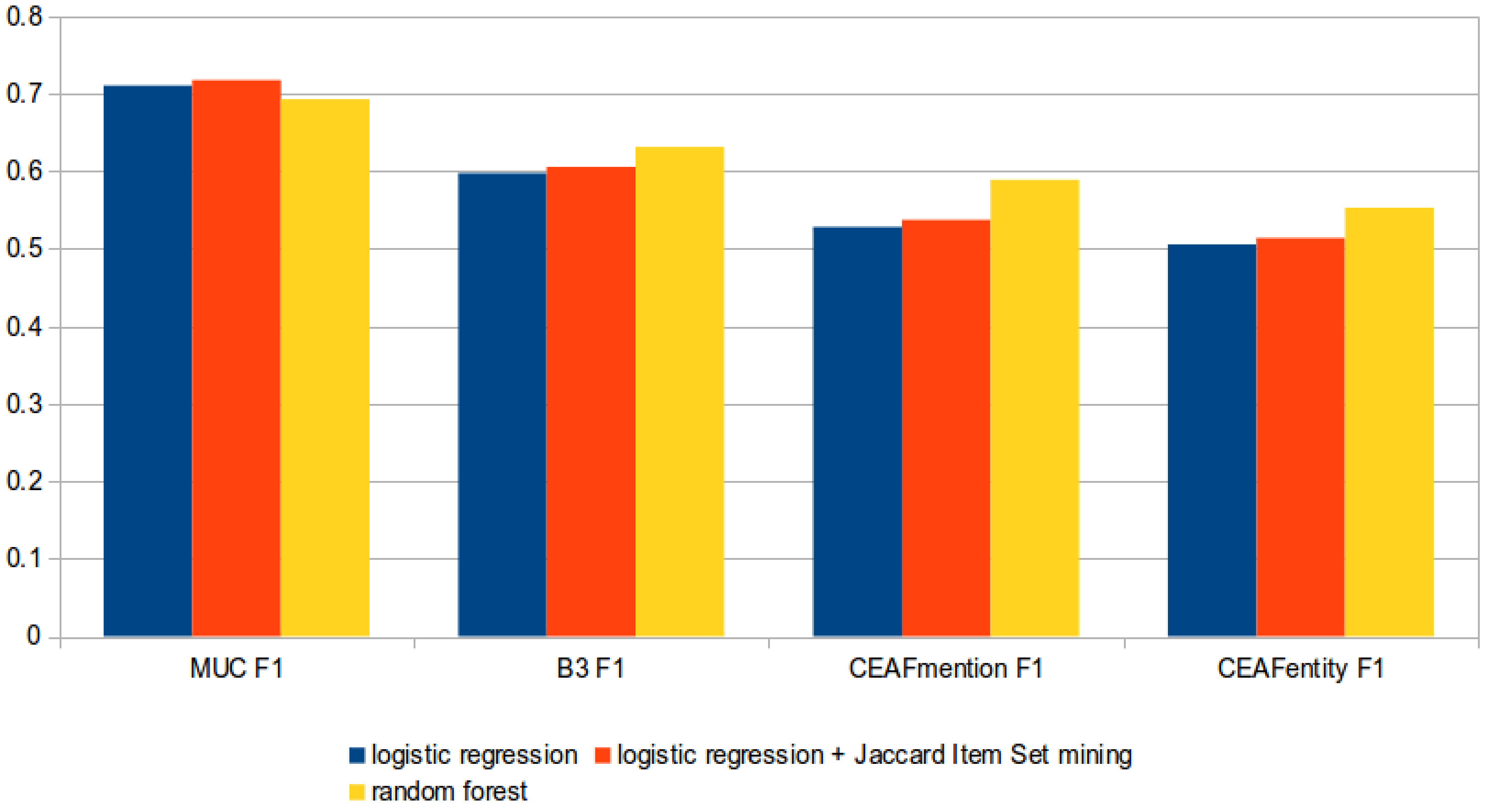
| | ANT_PER | ANAF_PER | LEX_SIM>0.5 | NUM | ANIM | SENT_ST |
|---|---|---|---|---|---|---|
| TRUE | F | F | T | F | F | T |
| FALSE | T | T | F | F | T | F |
| TRUE | T | F | F | F | T | T |
| FALSE | F | F | F | T | F | F |
| TRUE | T | F | T | F | T | T |
| FALSE | T | T | F | F | T | T |
| TRUE | F | T | T | F | F | F |

ANT_PER & ANAF_PER & ~(LEX_SIM>0.5)

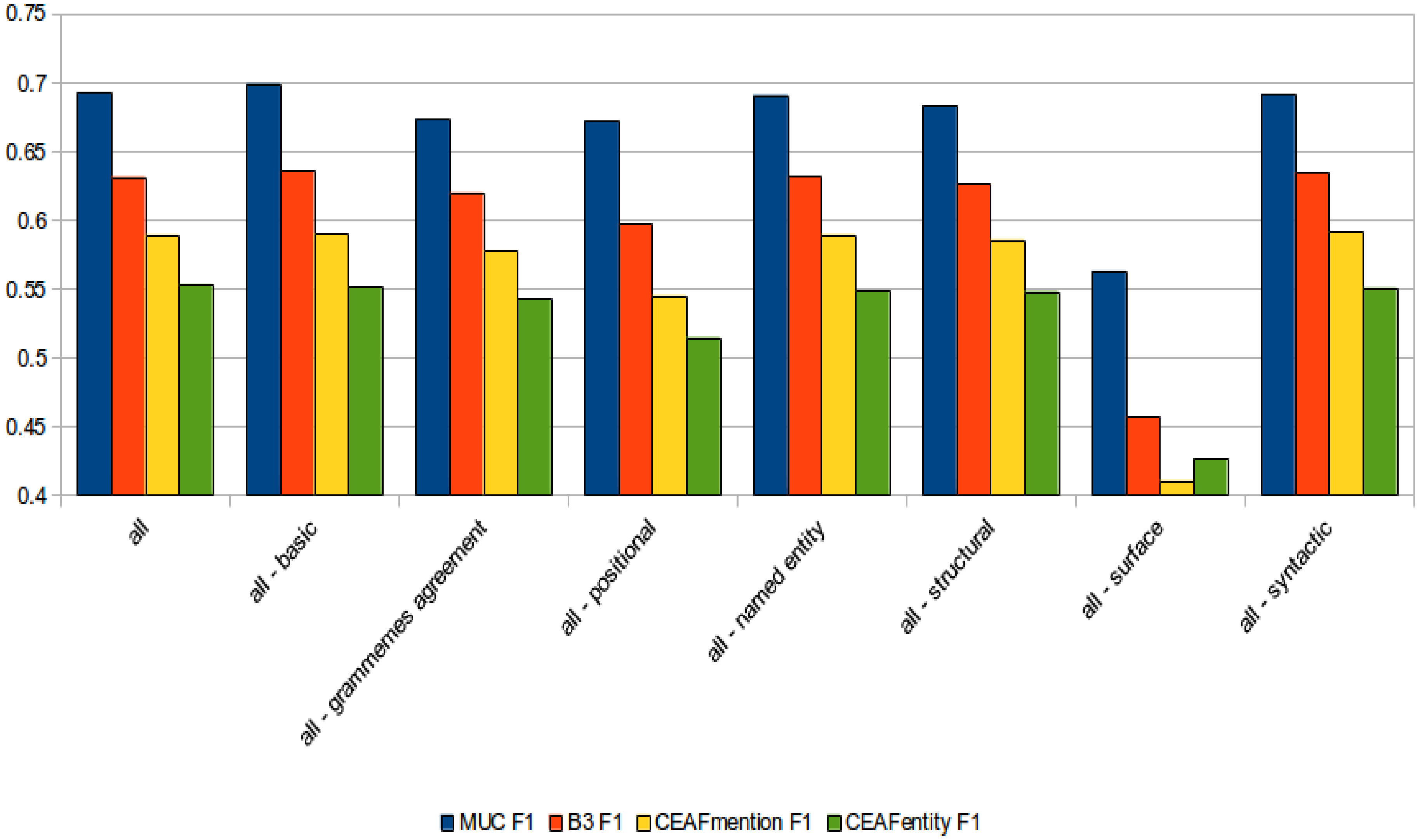
*Uryupina O., Moschitti A. (2015)*
*A State-of-the-Art Mention-Pair Model for Coreference Resolution*

*Segond M., Borgelt C. (2011)*
*Item Set Mining Based on Cover Similarity*

# Coreference resolution. Selecting classifier



# Coreference resolution. Ablation analysis



# Coreference resolution.
# In the wild

| | Precision | Recall | F1 |
|---|---|---|---|
| MUC | 0.4768 | 0.3741 | 0.4189 |
| B3 | 0.4104 | 0.2957 | 0.3431 |
| $\text{CEAF}_{\text{mention}}$ | 0.4024 | 0.3702 | 0.3854 |
| $\text{CEAF}_{\text{entity}}$ | 0.2525 | 0.3433 | 0.2906 |

# Future work

- Experiments with more machine learning algorithms and approaches.
- Using various clustering algorithms for word embeddings.
- Detailed analysis of features, assumed useless in ablation experiments.
- Tuning coreference resolution algorithm for different mention types.

# Credits

Alexandra Khadzhiiskaia

Ivan Andrianov

Thank you!