Aspect extraction from reviews using conditional random fields

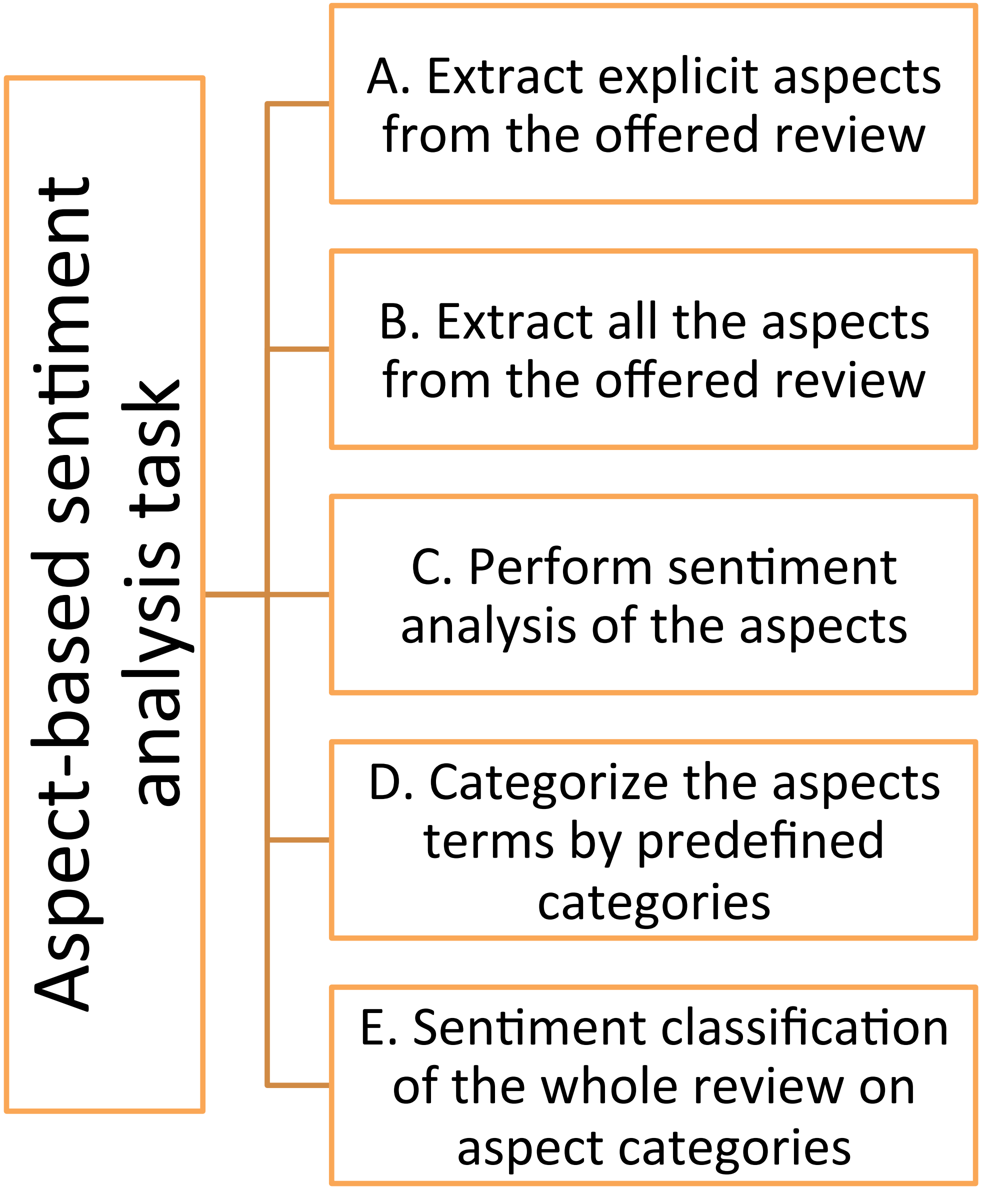
Yuliya Rubtsova

Sergey Koshelnikov

A.P. Ershov Institute of Informatics Systems
Russian Academy of Sciences
Siberian Branch

IIS SB RAS - 25 years!

SentiRuEval-2015



Aspect-based sentiment analysis task

A. Extract explicit aspects from the offered review

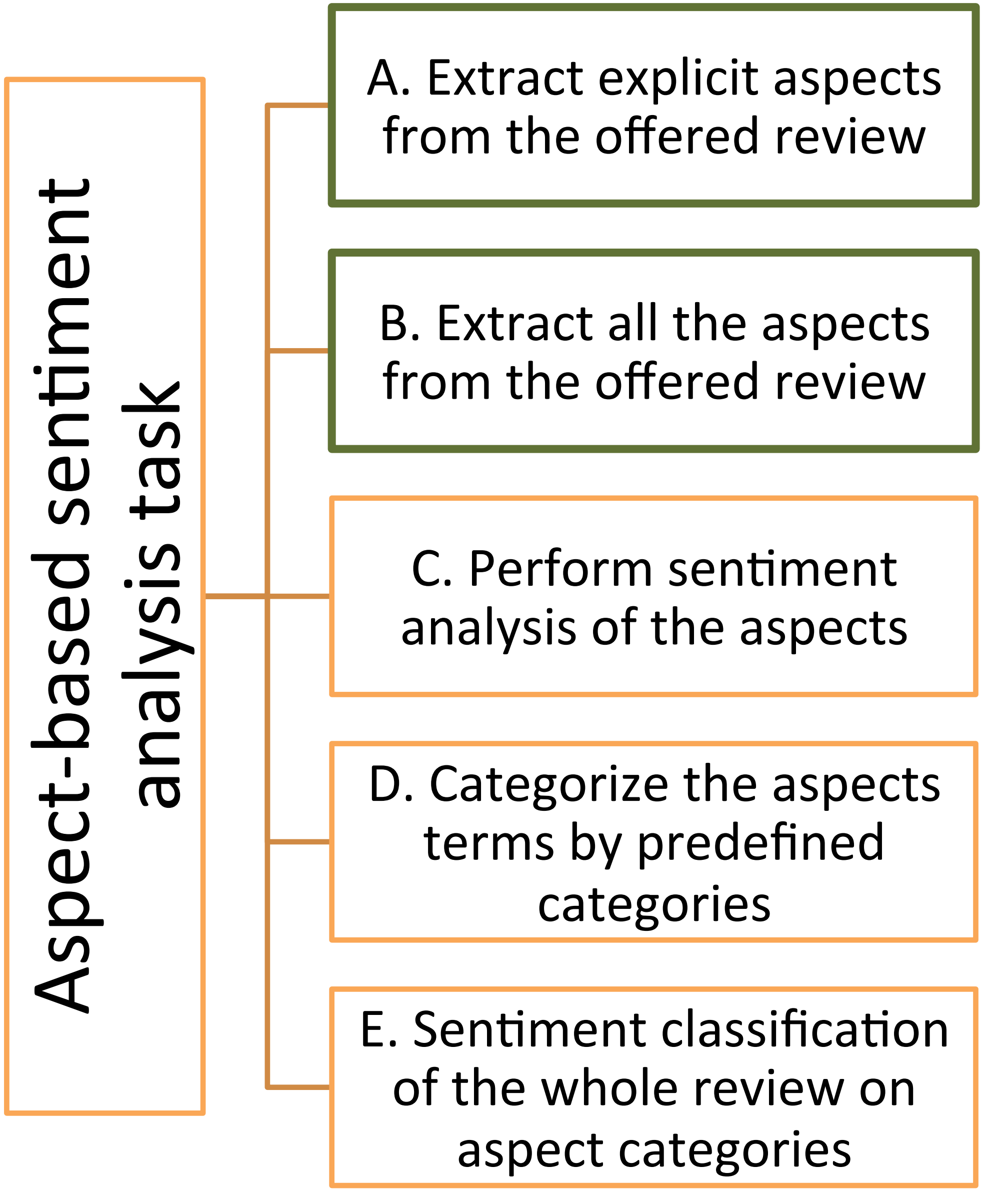
B. Extract all the aspects from the offered review

C. Perform sentiment analysis of the aspects

D. Categorize the aspects terms by predefined categories

E. Sentiment classification of the whole review on aspect categories

SentiRuEval-2015



Aspect-based sentiment analysis task

A. Extract explicit aspects from the offered review

B. Extract all the aspects from the offered review

C. Perform sentiment analysis of the aspects

D. Categorize the aspects terms by predefined categories

E. Sentiment classification of the whole review on aspect categories

Major approaches to extract aspects

Frequency of nouns and/or noun phrases (Hu and Liu, 2004)

Simultaneous extraction of both sentiment words (user opinions) and aspects

Supervised machine learning (HMM, Jin et al., 2009 and CRF, Jakob and Gurevych, 2010).

Unsupervised machine learning or topic modeling (Titov and McDonald, 2008; Brody and Elhadad, 2010)

Major approaches to extract aspects

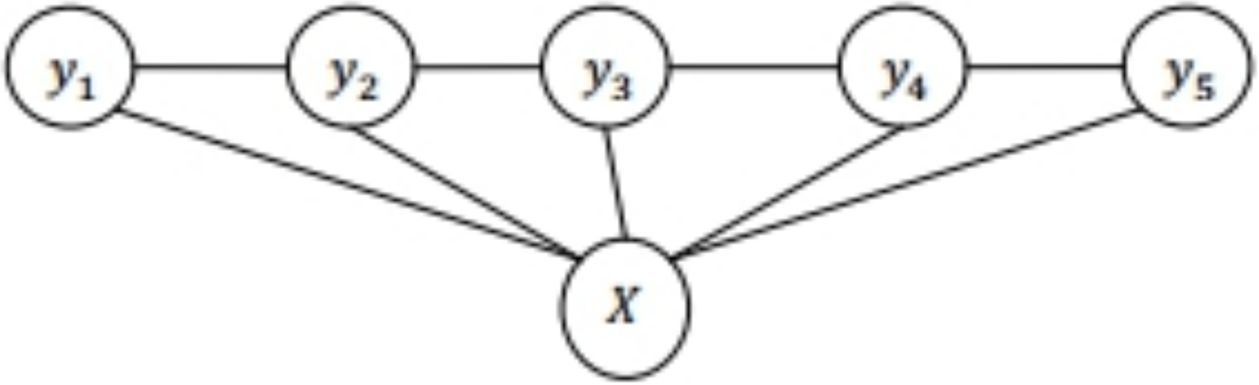
Frequency of nouns and/or noun phrases (Hu and Liu, 2004)

Simultaneous extraction of both sentiment words (user opinions) and aspects

Supervised machine learning (HMM, Jin et al., 2009 and CRF, Jakob and Gurevych, 2010).

Unsupervised machine learning or topic modeling (Titov and McDonald, 2008; Brody and Elhadad, 2010)

Conditional Random fields (CRF)



CRFs are a type of discriminative undirected probabilistic graphical model. It is used to encode known relationships between observations and construct consistent interpretations.

Conditional Random fields (CRF)

Let G be a graph such that Y = $(Y_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field when the random variables Yv, conditioned on X, obey the Markov property with respect to the graph

$$P(y_v \mid Y_{V \backslash \{v\}}, X) = P(y_v \mid Y_{o(v)}, X),$$

Conditional Random fields (CRF)

$$P(Y \mid X) = \frac{1}{\mathrm{Z}(X)} \exp(\sum\nolimits_{c \in C} \lambda_c f_c(y_c, X)),$$

Where Z(x) is normalization factor,
C – set of all graphs' cliques,
f_c – set of features,
λi – factors.

CRF advantages

Relaxation of the independence assumptions

CRFs avoid the label bias problem

System description

Pre-processing

“s-e” – start of an explicit aspect term,
“c-e” – continuation of an explicit aspect term,
“s-i” – start of an implicit aspect term,
“c-i” – continuation of an implicit aspect term,
“s-f” – start of an implicit aspect term,
“c-f” – continuation of an implicit aspect term, “O” indicates not an aspect term.

System description

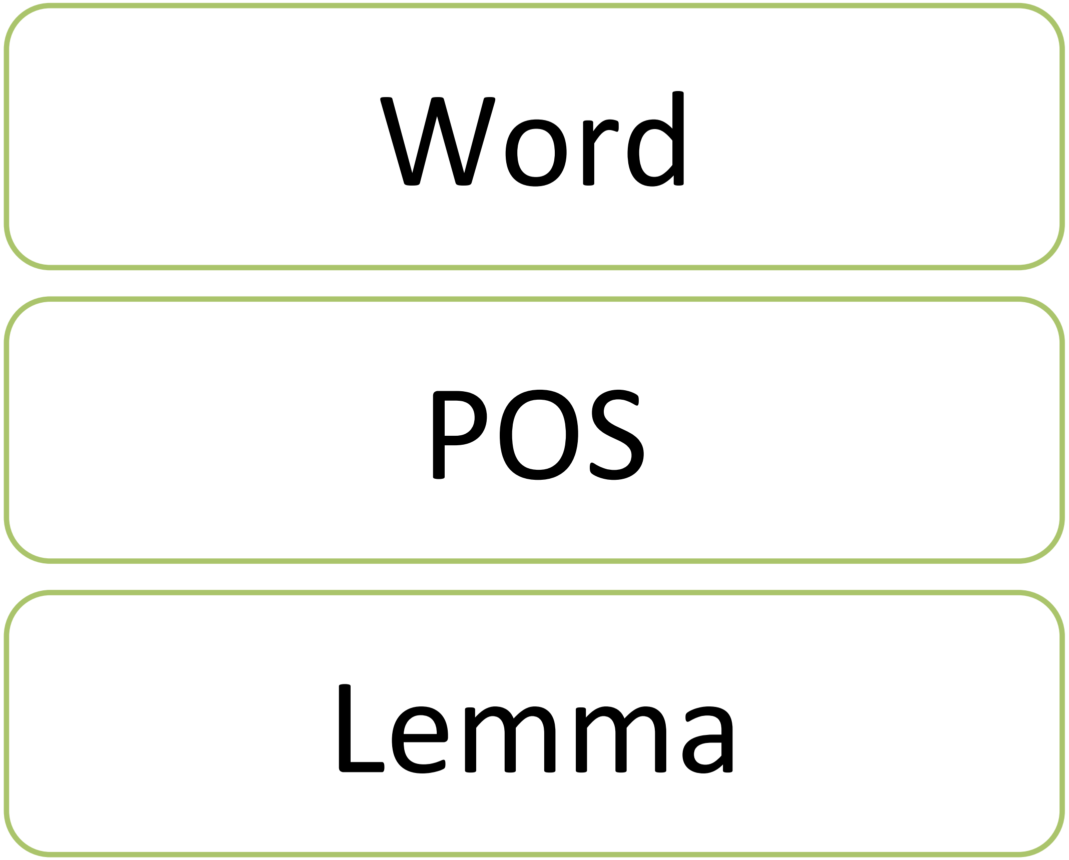
Pre-processing

To extract syntactic features (e.g. POS, lemma) we used TreeTagger for Russian (Sharoff, 2008)

We also converted all the capital letters into lowercase

System description

features



Word

POS

Lemma

System description

example

Очень дружелюбное место, с порога встречают симпатичные работники, тёплый, уютный интерьер и зажигательная музыка

Very friendly place where pretty staff meet from the threshold, warm and cozy interior and incendiary music

System description

example

w[0]=**очень** w[-1]=null w[1]=дружелюбное pos[0]=r O

w[0]=**дружелюбное** w[-1]=очень w[1]=место pos[0]=a O

w[0]=**место** w[-1]=дружелюбное w[1]=null pos[0]=n s-e

w[0]=**с** w[-1]=null w[1]=порога pos[0]=s O

w[0]=**порога** w[-1]=с w[1]=встречают pos[0]=n O

w[0]=**встречают** w[-1]=порога w[1]=симпатичные pos[0]=v s-e

w[0]=**симпатичные** w[-1]=встречают w[1]=работники pos[0]=a O

w[0]=**работники** w[-1]=симпатичные w[1]=тёплый pos[0]=n O

w[0]=**тёплый** w[-1]=работники w[1]=уютный pos[0]=a O

w[0]=**уютный** w[-1]=тёплый w[1]=интерьер pos[0]=a O

w[0]=**интерьер** w[-1]=уютный w[1]=и pos[0]=n s-e

w[0]=**и** w[-1]=интерьер w[1]=зажигательная pos[0]=c O

w[0]=**зажигательная** w[-1]=и w[1]=музыка pos[0]=a O

w[0]=**музыка** w[-1]=зажигательная w[1]=null pos[0]=n s-e

System description

System 1: CRF with all the above-mentioned labels. We used s-e, c-e and O labels for explicit aspect extraction to perform Task A and s-e, c-e, s-i, c-i, s-f, c-f, O to extract all the aspects for Task B.

System 2: Combination of the results of two CRFs —CRF for extraction of explicit aspect terms and CRF for extraction of implicit aspect terms + sentiment facts terms (not explicit).

Task A was performed using System 1 and Task B — using both systems.

Results

F-measure

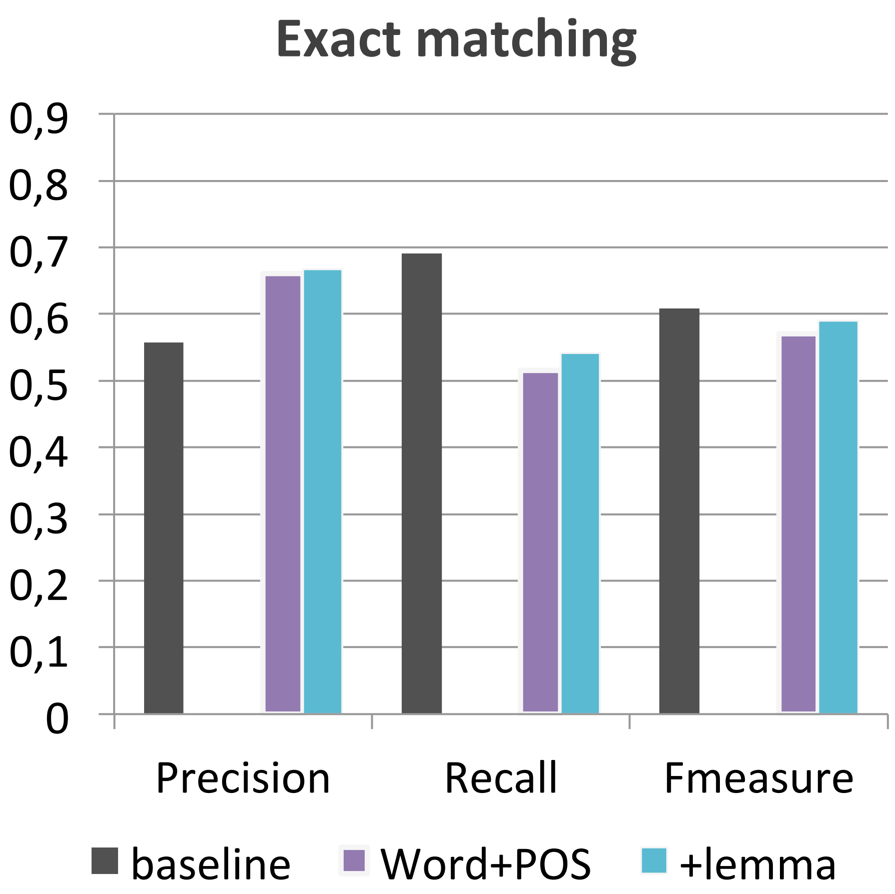
Exact matching and partial matching.

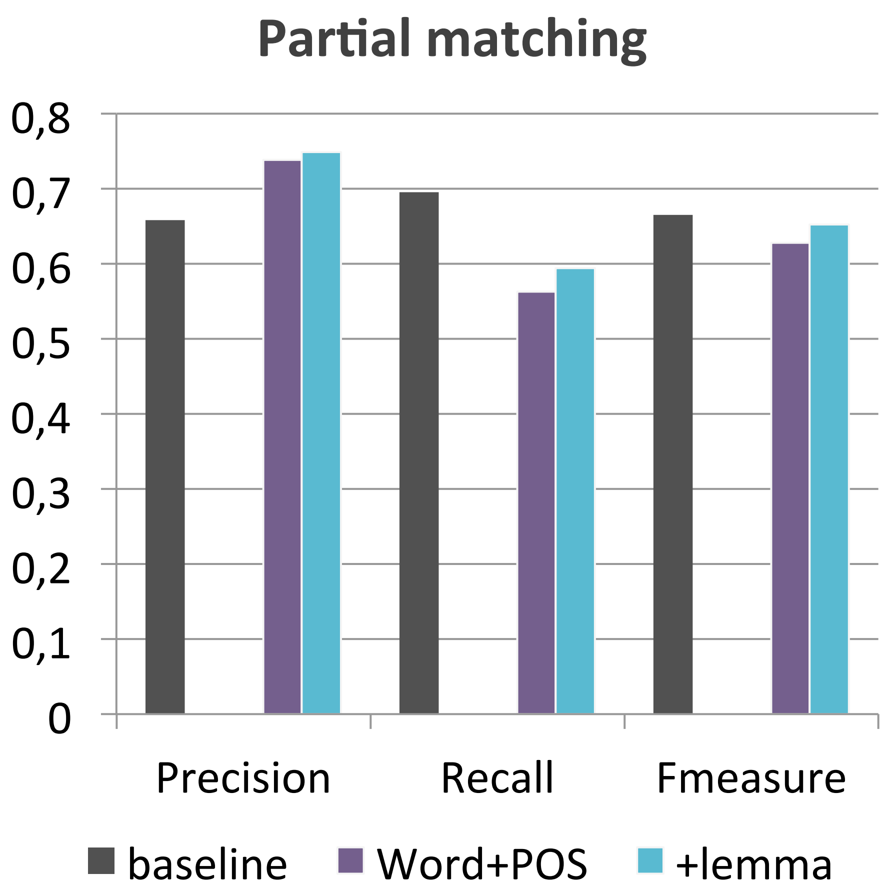
Macro F1-measure means in this case calculating F1-measure for every review and averaging the obtained values.

Micro F – partial matching, the intersection between gold standard and extracted term was calculated.

Results

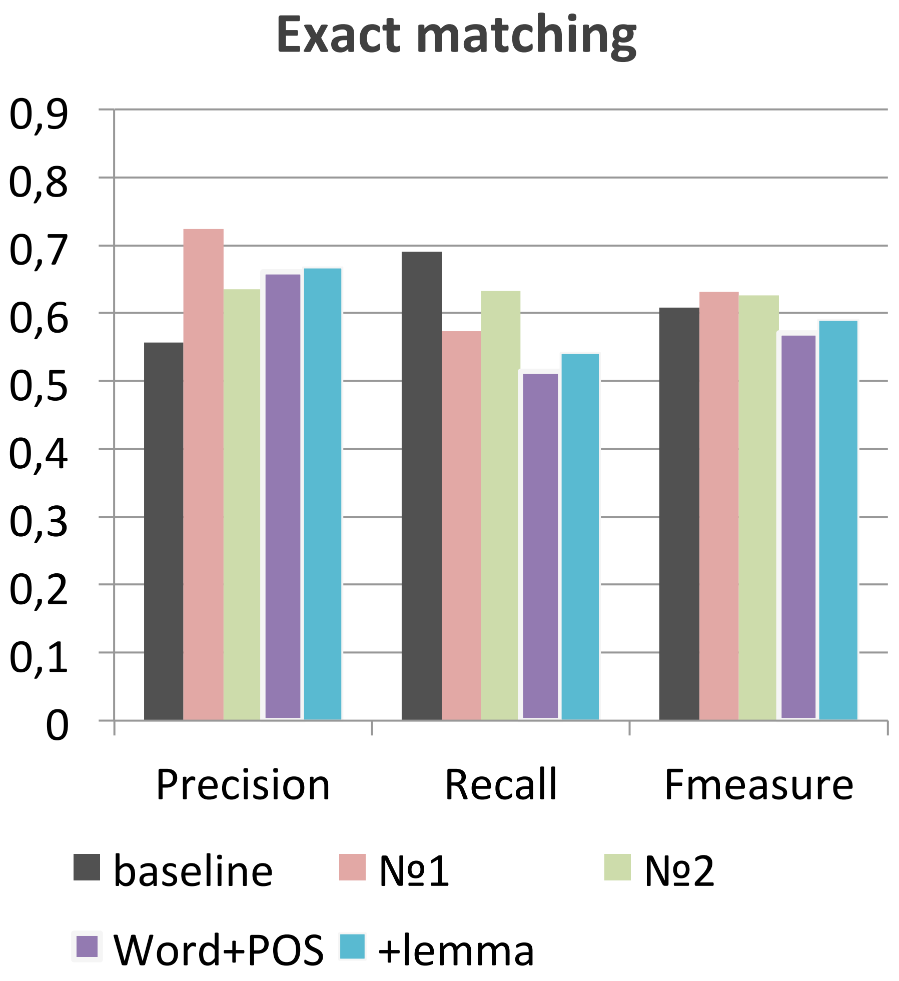
Task A restaurant domain in comparison to baseline

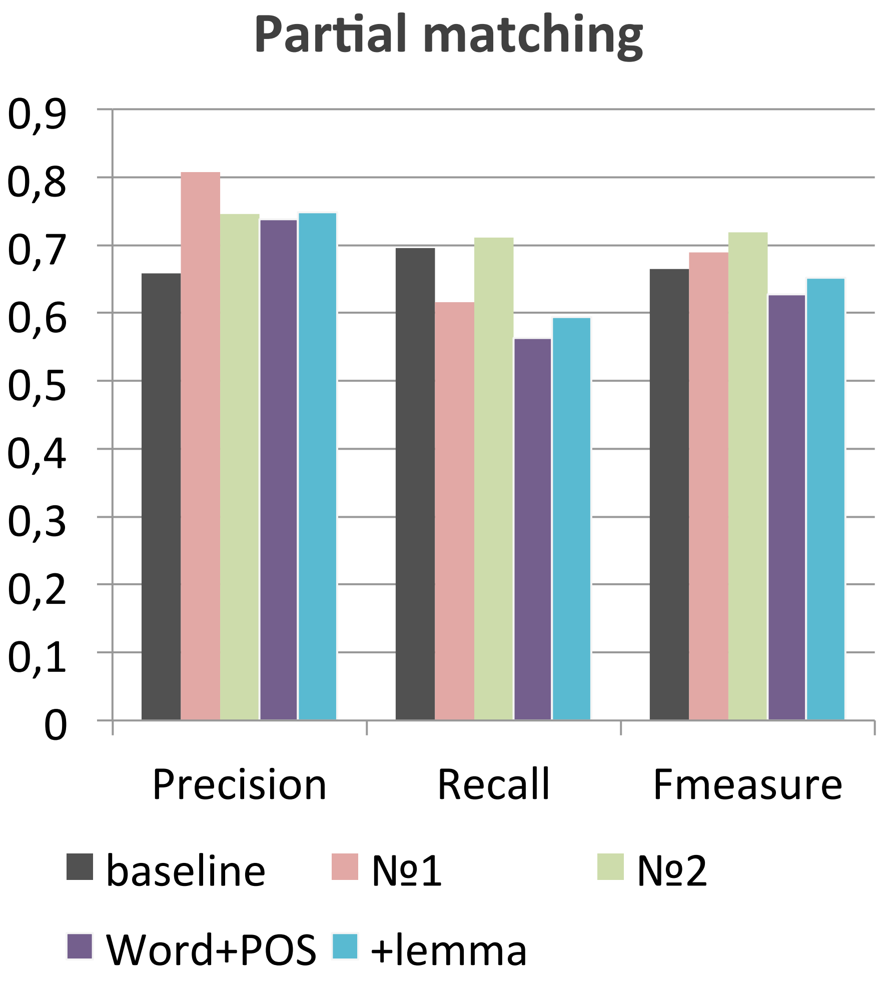




Results

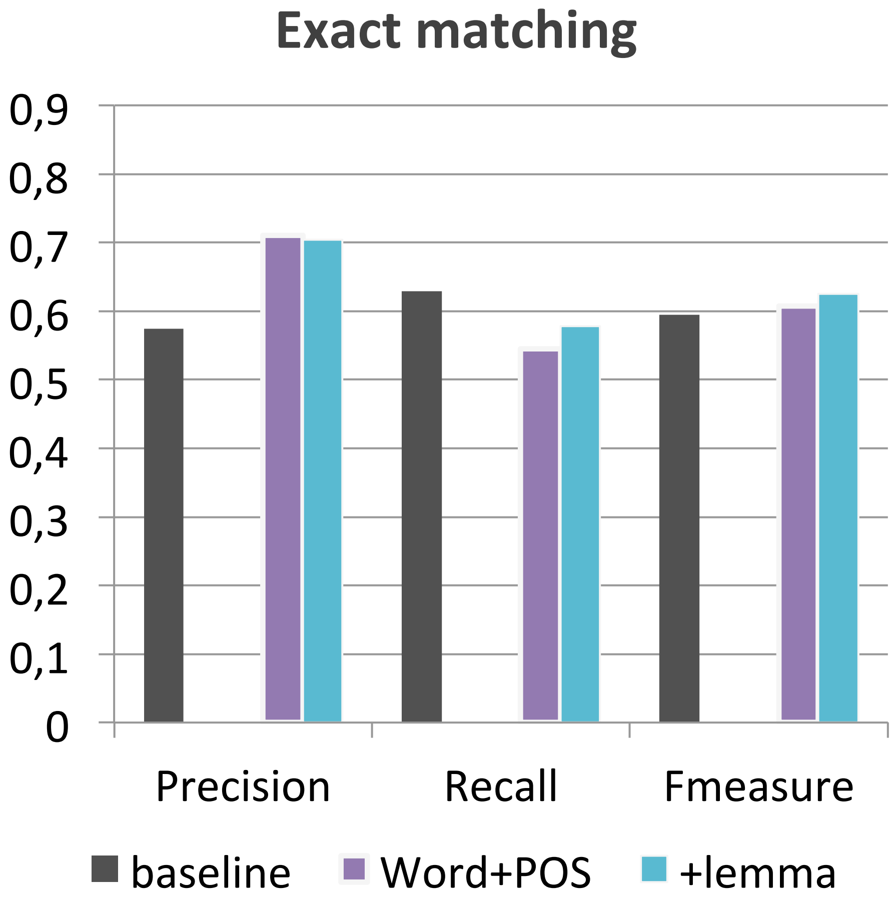
Task A restaurant domain in comparison to the best results



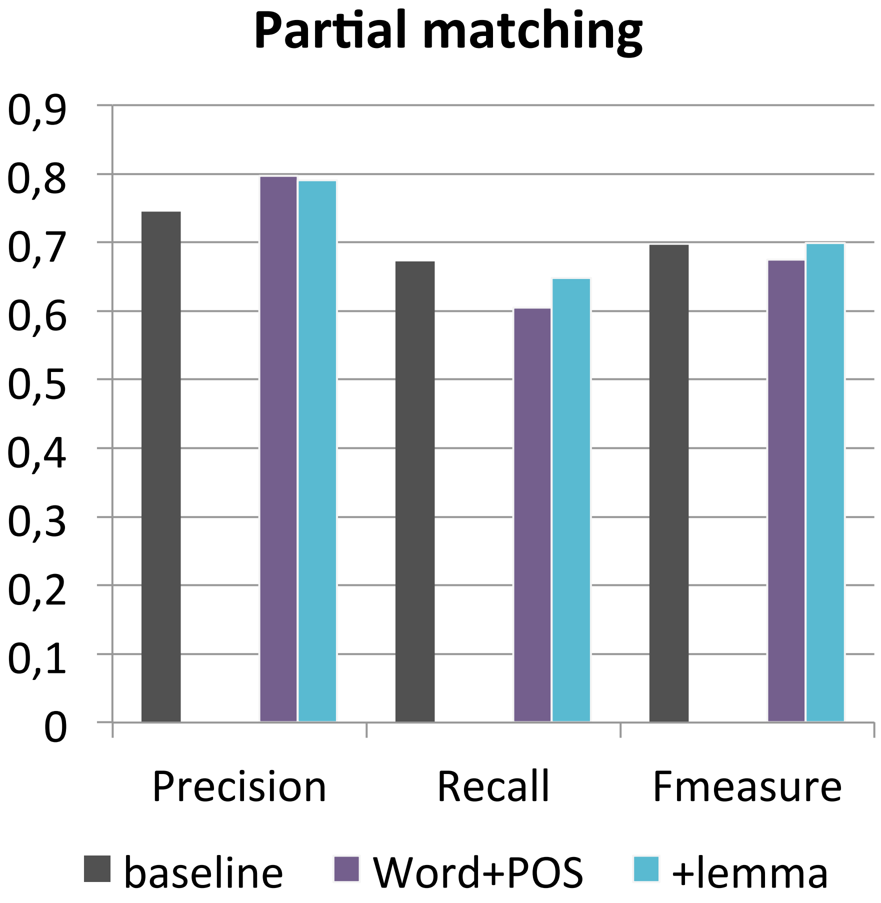


Results

Task A car domain in comparison to baseline



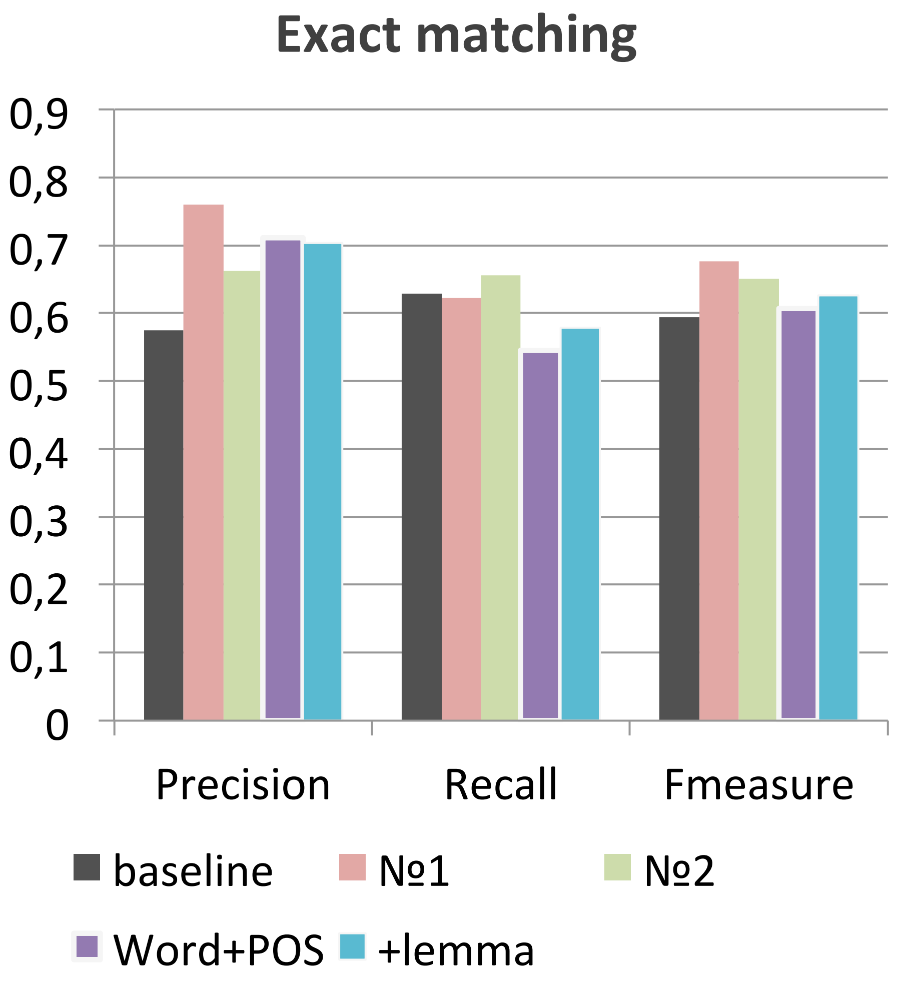
Exact matching

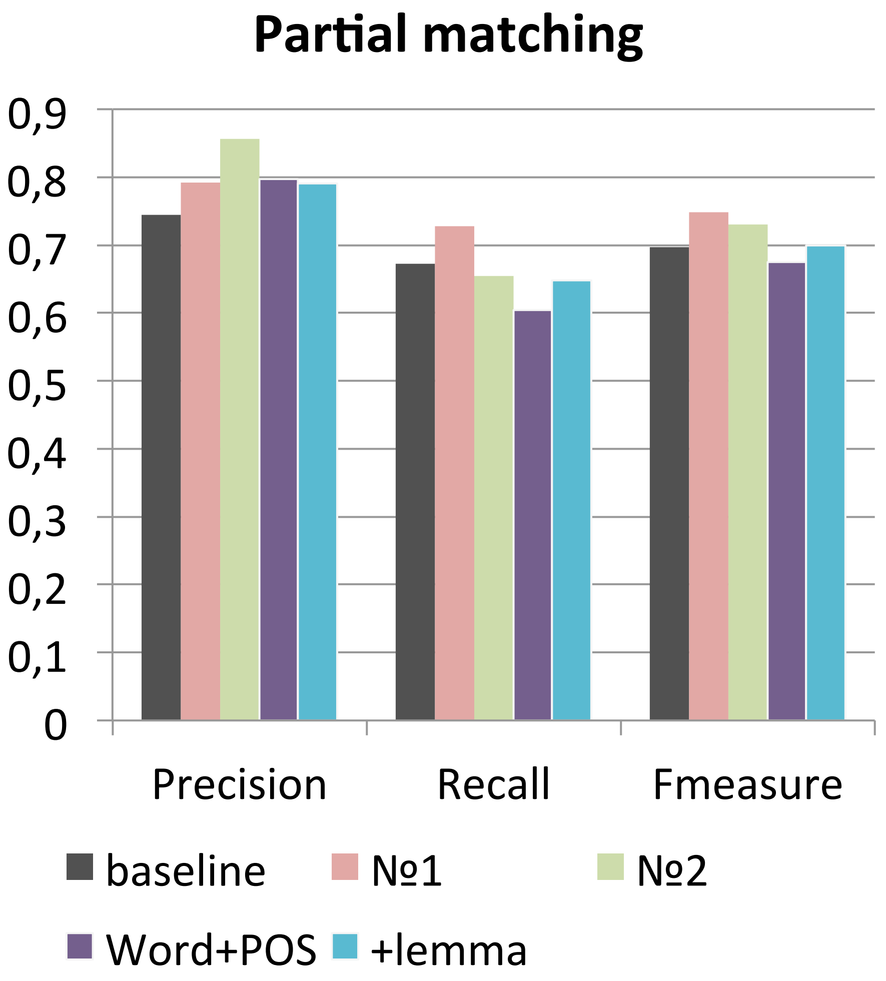


Partial matching

Results

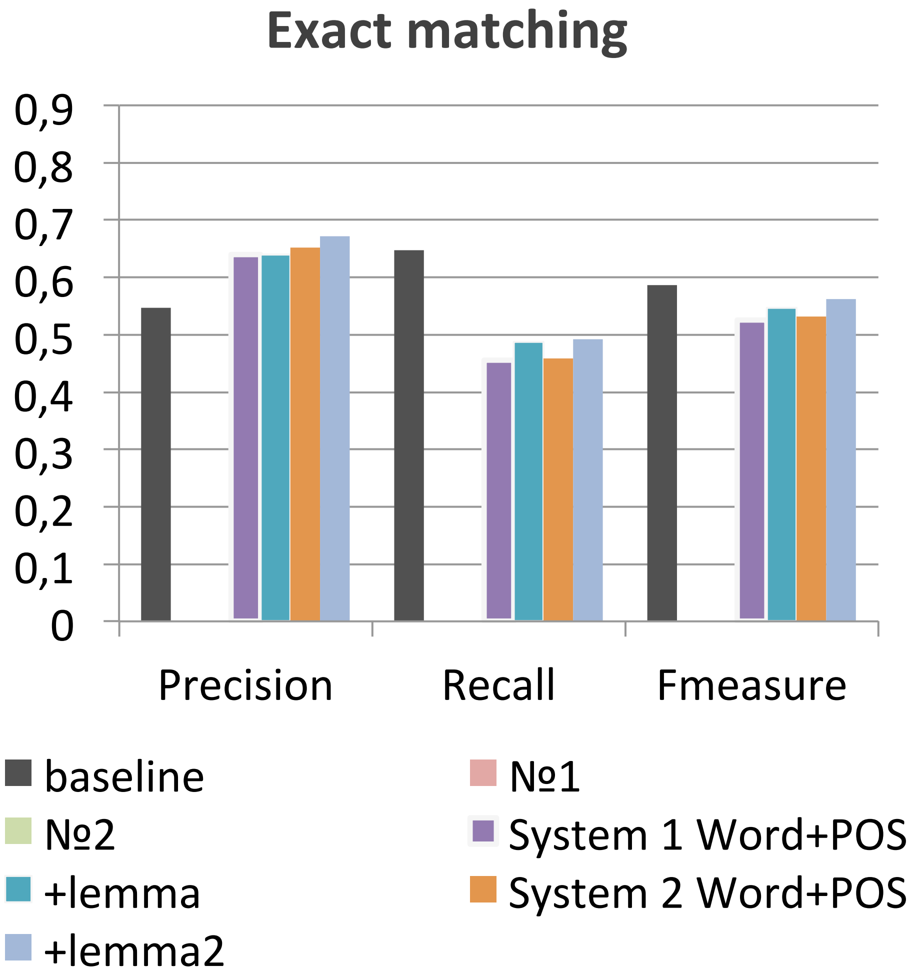
Task A car domain in comparison to the best results

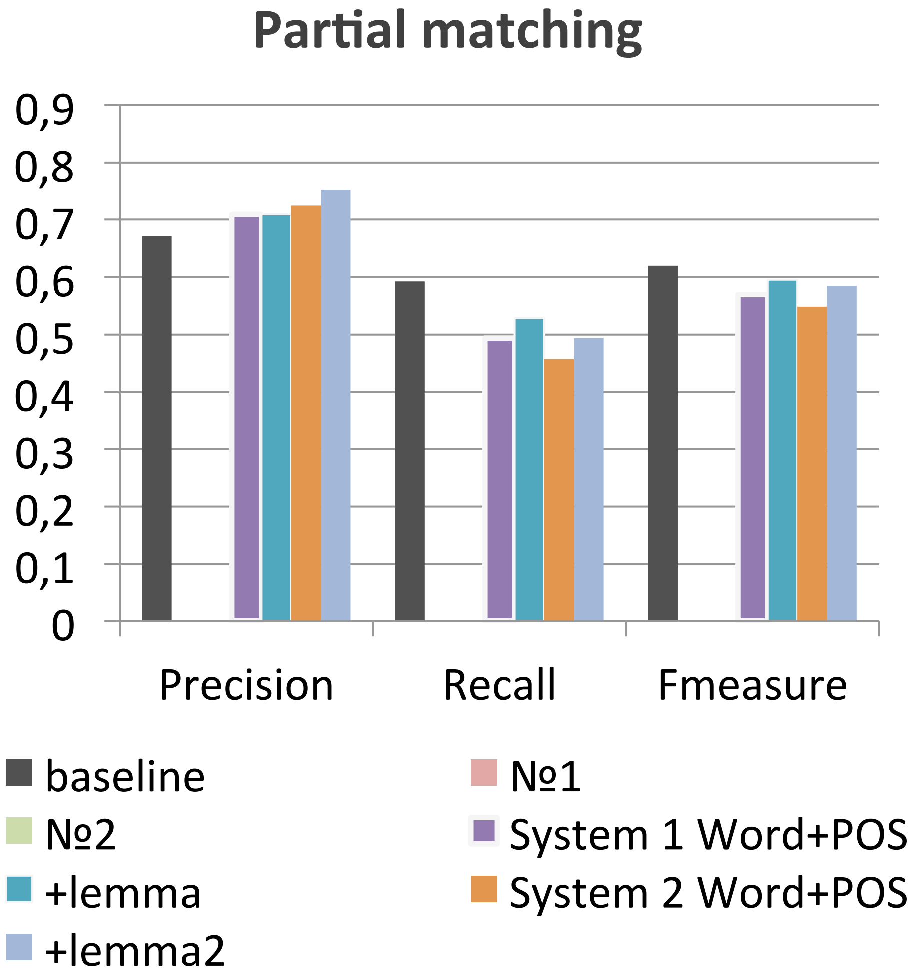




Results

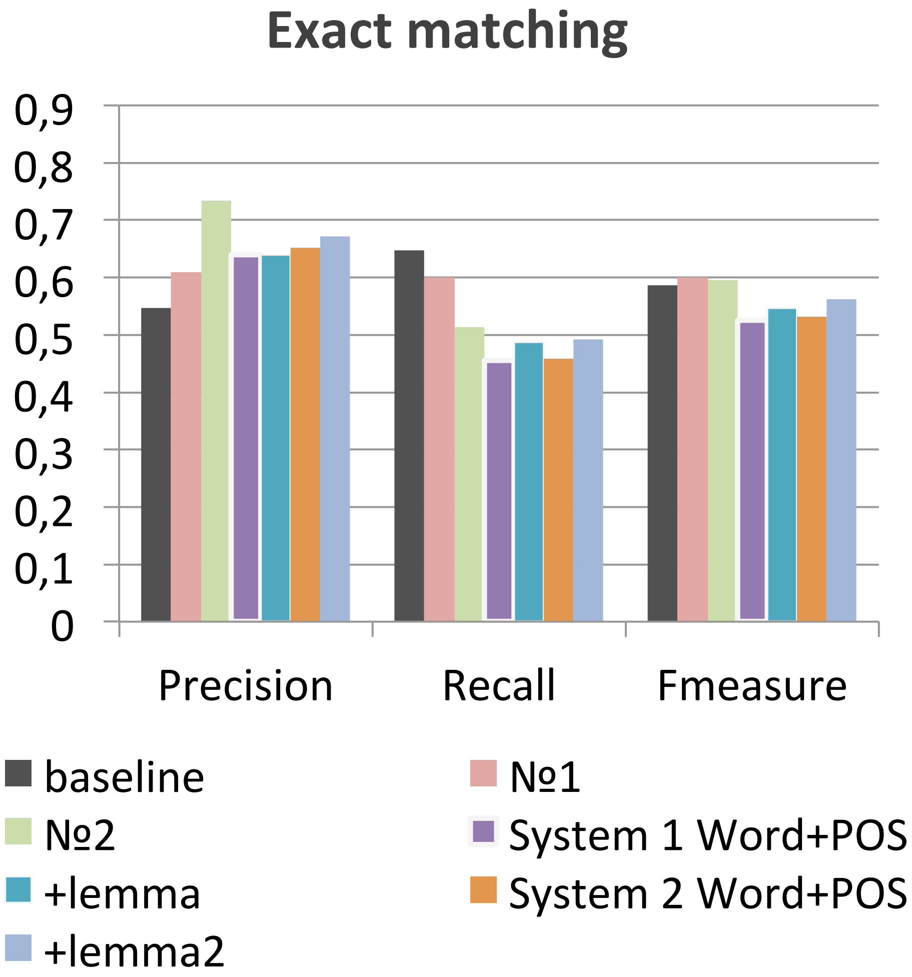
Task B restaurant domain in comparison to baseline

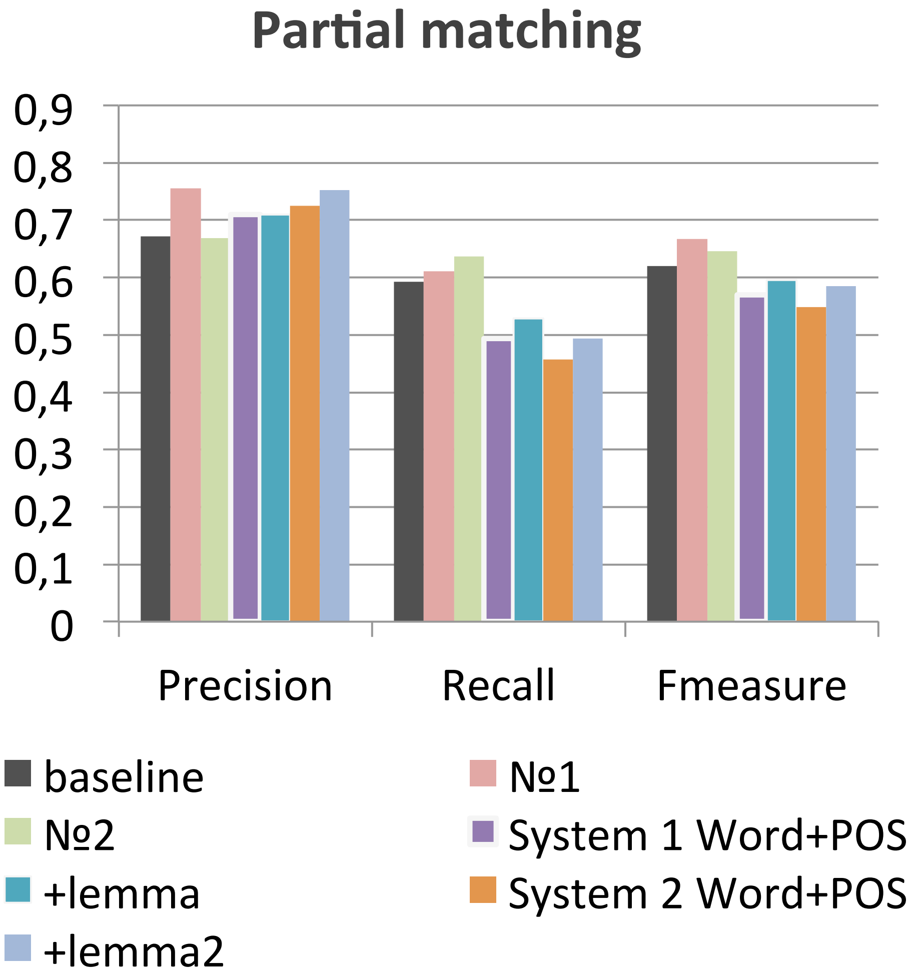




Results

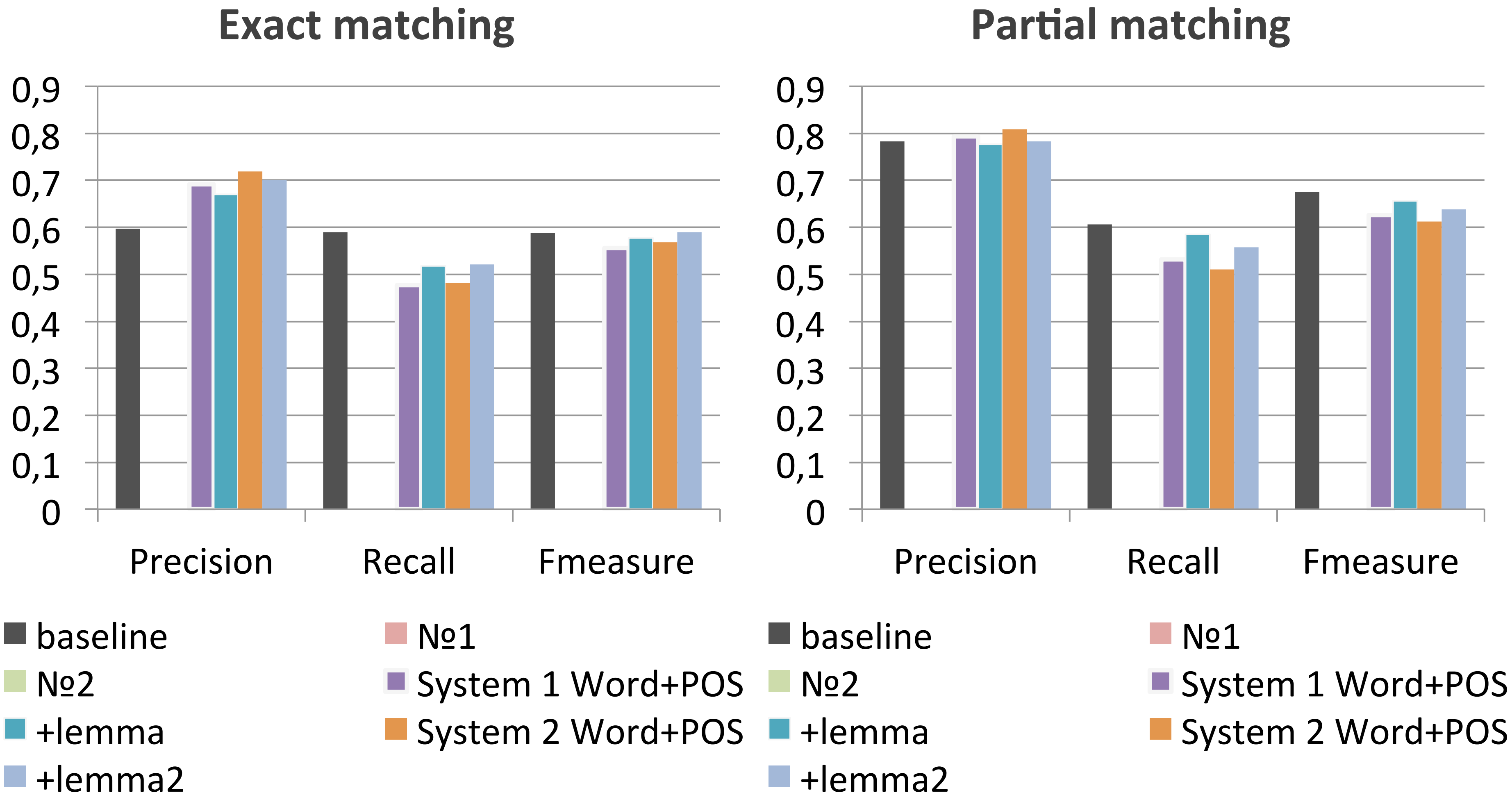
Task B restaurant domain in comparison to the best results





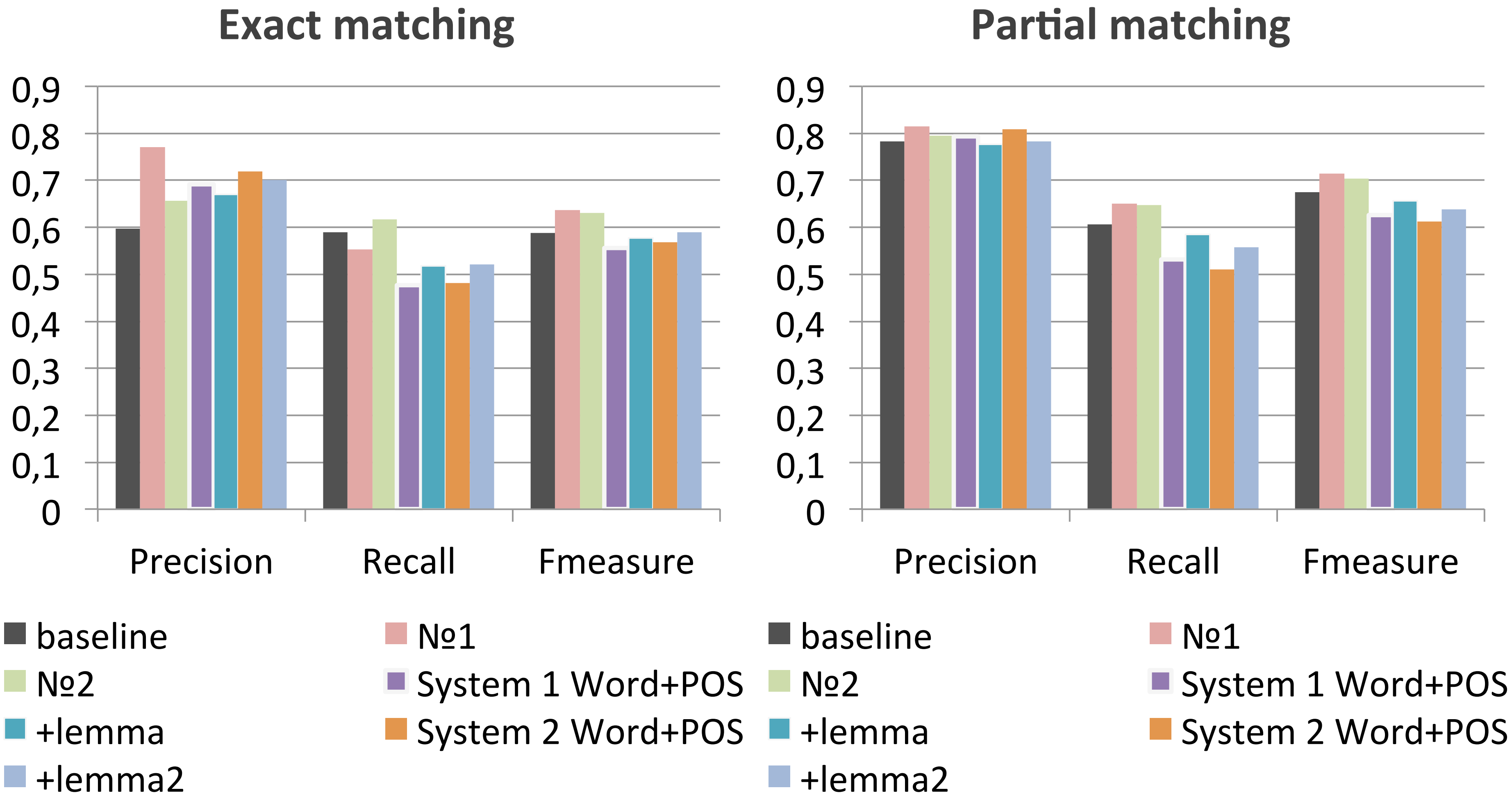
Results

Task B car domain in comparison to baseline



Results

Task B car domain in comparison to the best results



Error Analysis

Not recognized

excessively recognized

Error distribution

	Restaurants	**Car**
Not recognized	67,1%	63%
excessively recognized	32,9%	37%

Error types

1. Technical errors

1.1 Special symbols:

Etalon: Салат &quot;цезарь&quot;

System: Салат &quot;цезарь

1.2 Lower case:

Can’t recognize ie “TO” (technical maintenance in car domain) and “то” (the particle)

Error types

2. Not recognized

2.1 Shortness

Рублей –> руб. –> р. (rubles -> rub -> R.)

2.2 listings

Овощи, **салаты «Цезарь»**, лосось (Vegetables, salads "Caesar", salmon)

Error types

3. Partly recognition

3.1. Before head word

“Добавляла **вина**” (pour **wine**)

“Официант **хамил**” (The waiter **was rude**)

3.2. After head word

“**местечко** в углу” (**a place** in the corner)

4. Excessively recognized

4.1 Not always good deal with named entities

Александр (Alexander)

Conclusion

- Even a small features for CRF demonstrates quite a good performance. The results of our systems was comparable to the best results of SentiRuEval participants.
- Subsequently we are going to add statistical methods as a CRF feature.

A.P. Ershov Institute of Informatics Systems
Russian Academy of Sciences
Siberian Branch

IIS SB RAS - 25 years!

Thank you!

Yuliya Rubtsova
yu.rubtsova@gmail.com
study.mokoron.com