Named Entity Normalization for Fact Extraction Task

Popov A. M.,
Adaskina Yu. V., Andreyeva D. A.,
Charabet Ja. K., Moskvina A. D.,
Protopopova E. V., Yushina T. A.

About Us

- We are post graduate students from SPSU and colleagues working in the field;
- Project started in October 2015;
- At the beginning we had our own tokenizer and morphological analyzer;
- But we decided to participate in all three FactRuEval tasks.

About FactRuEval

FactRuEval – an independent competition for information extraction systems for Russian.

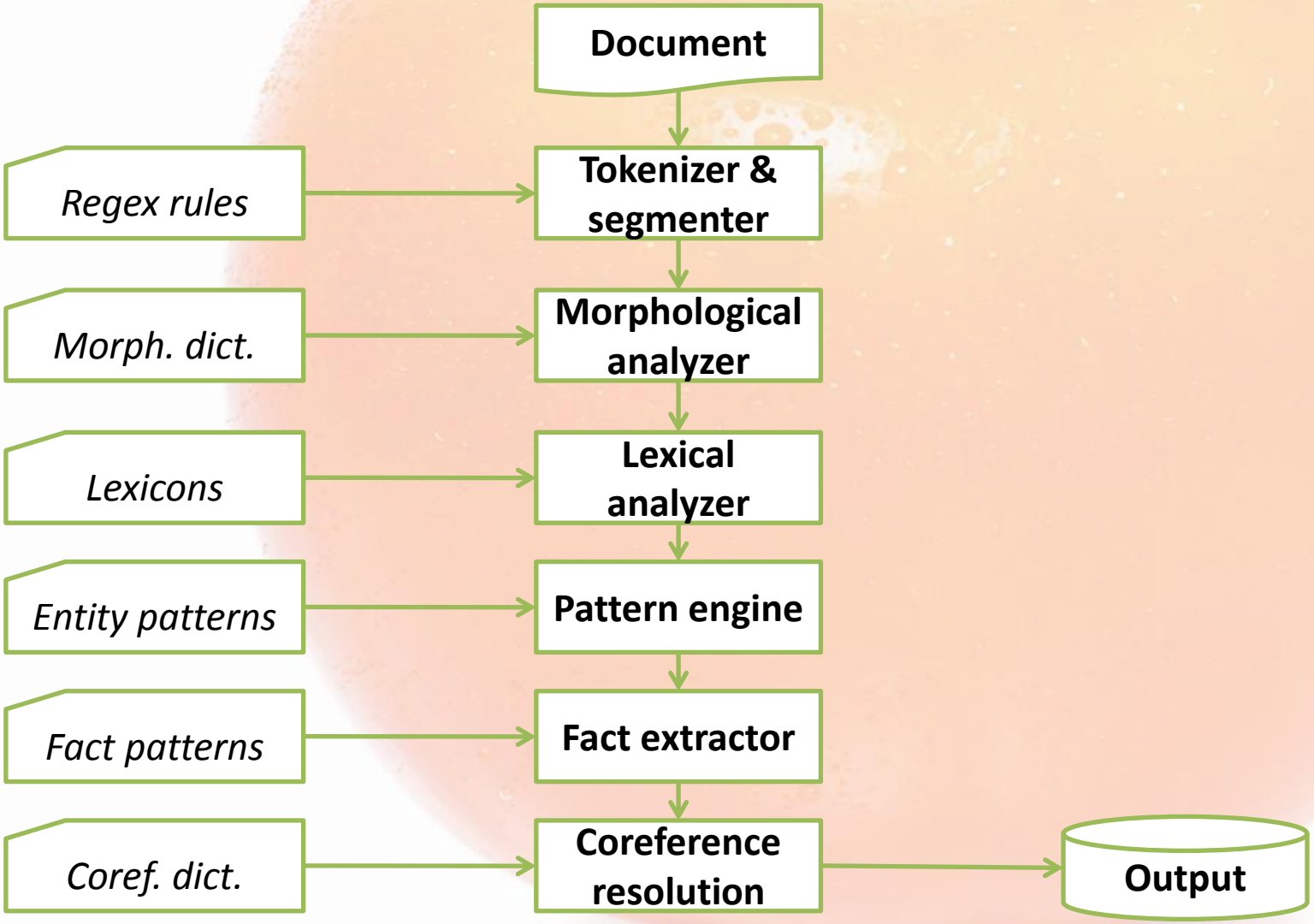
FactRuEval 2016 includes three tasks:

- Named Entity Extraction;
- Named Entity Normalization;
- Fact Extraction.

Our Approach

- Rule-based;
- Classical pipeline system;
- Not relying on syntax;
- Context-free token-based regex-style rules for entity and fact extraction.

Our Workflow



Named Entity Recognition

- For entity extraction we used a set of rules describing token sequences in terms of lexical and grammatical restrictions;
- Rules are written as JSON-documents.

Entity Extraction Rule Example (1)

This rule *matches*: “следственный комитет МВД”, “нефтяная компания Total”, etc.

```
"OrganizationAdjKeyOrgname": {
    "ngram": [
        { "tag": "Adj", "field": ["Name"], "norm": ["nom", "$3"] },
        { "tag": "OrgKey", "field": ["Name"] },
        { "tag": "WOrgName|org", "field": ["Name"], "norm": ["$SELF"] },
    ],
    "agreements": [
        { "Case": [1, 2] },
        { "Number": [1, 2] },
        { "Gender": [1, 2] },
    ],
    "entity": {
        "type": "ORG",
        "tag": ["ECompany", "Noun"],
        "head": 2,
        "priority": 1,
    },
},
```

Entity Extraction Rule Example (2)

Here we describe token sequence to be found

```
"OrganizationAdjKeyOrgname": {
    "ngram": [
        { "tag": "Adj", "field": ["Name"], "norm": ["nom", "$3"] },
        { "tag": "OrgKey", "field": ["Name"] },
        { "tag": "WOrgName|org", "field": ["Name"], "norm": ["$SELF"] },
    ],
    "agreements": [
        { "Case": [1, 2] },
        { "Number": [1, 2] },
        { "Gender": [1, 2] },
    ],
    "entity": {
        "type": "ORG",
        "tag": ["ECompany", "Noun"],
        "head": 2,
        "priority": 1,
    },
},
```

Entity Extraction Rule Example (3)

```
"OrganizationAdjKeyOrgname": {
    "ngram": [
        { "tag": "Adj", "field": ["Name"], "norm": ["nom", "$3"] },
        { "tag": "OrgKey", "field": ["Name"] },
        { "tag": "WOrgName|org", "field": ["Name"], "norm": ["$SELF"] },
    ],
    "agreements": [
        { "Case": [1, 2] },
        { "Number": [1, 2] },
        { "Gender": [1, 2] },
    ],
    "entity": {
        "type": "ORG",
        "tag": ["ECompany", "Noun"],
        "head": 2,
        "priority": 1,
    },
},
```

Here we restrict the sequence with grammatical agreements

Entity Extraction Rule Example (4)

```
"OrganizationAdjKeyOrgname": {
    "ngram": [
        { "tag": "Adj", "field": ["Name"], "norm": ["nom", "$3"] },
        { "tag": "OrgKey", "field": ["Name"] },
        { "tag": "WOrgName|org", "field": ["Name"], "norm": ["$SELF"] },
    ],
    "agreements": [
        { "Case": [1, 2] },
        { "Number": [1, 2] },
        { "Gender": [1, 2] },
    ],
    "entity": {
        "type": "ORG",
        "tag": ["ECompany", "Noun"],
        "head": 2,
        "priority": 1,
    },
},
```

And here we describe the extracted entity

Entity Normalization (1)

- We perform normalization of extracted entities by means of special notes in extraction rules called “normalization directives”.
- Each normalization directive tells the system how to normalize a specific element of the rule matched to the token sequence.

Entity Normalization (2)

We have the following set of directives implemented:

- Default – word is normalized to the dictionary lemma;
- Self – word is left in the exact form it occurred in the text;
- Explicit – we can describe a set of grammemes of the desired wordform;
- Implicit – we can describe the target grammemes of the desired wordform as an agreement;
- Conditional – we can add some grammemes to the desired wordform if and only if these grammemes are present in the tag of the source wordform.

Entity Normalization (3)

```
{ "tag": "Adj", "field": ["Name"], "norm": ["nom", "$3"] },
{ "tag": "OrgKey", "field": ["Name"] },
{ "tag": "WOrgName|org", "field": ["Name"], "norm": ["$SELF"] },
```

Normalization directives are highlighted:

- “nom” means to normalize word to Nominative Case
- “$3” means to keep grammems of the third agreement (gender)
- $SELF means to leave the wordform untouched
- There is no explicit directive for second element, thus the word will be normalized to dictionary lemma

Fact Extraction (1)

- Since we used no syntactic information;

and

- Entities filling fact fields are often non-continual (i.e. n-grams would not work);

we decided to accommodate

- A skip-gram rule engine for fact extraction.

Fact Extraction (2)

Example:

“Ранее, в январе этого года, суд запретил *консорциуму "Инвестиционно-металлургический союз"* осуществлять любые действия по отчуждению *принадлежащих* ему акций *"Криворожстали"*.”

Fact Extraction Rule Example (1)

This rule *matches*: “Посол Японии в России Масахару Коно будет снят с должности”

```
"Occupation2": {
    "skipgram": [
        { "tag": "EPosition|PersonPosition", "field": "Job" },
        { "tag": "ECompany|ELocation", "field": "Where" },
        { "tag": "EPerson", "field": "Person" },
    ],
    "ranges": [
        { "range": [1, 3], "max": 5 },
    ],
    "fact": {
        "priority": 1,
        "type": "Occupation",
        "head": 1,
        "restrict": [3],
    },
},
```

Fact Extraction Rule Example (2)

This result would be

```
---- #1 ----
Quality:
    Argument extraction quality = 1.00
    Identification quality = 1.00
    OVERALL = 1.00

STANDARD:
    [ occupation | who : коно масахару | position : Посол Японии в России |
    Посол | where : страна восходящего солнца | япония ]
TEST:
    [ occupation | where : япония | position : посол | who : коно масахару ]

ARGUMENTS:
    where : страна восходящего солнца | япония = where : япония
    who : коно масахару = who : коно масахару
    position : Посол Японии в России | Посол = position : посол
```

Entity Recognition Evaluation (1)

- Nearly 90% of locations are single-word entities and they are recognized by a morphological tag;
- Persons have limited possible combinations of Name, Surname and Patronymic, which are tagged morphologically, so the challenge here – unknown names and surnames;
- Organizations are the most difficult type, since they can only be described with high number of context patterns.

Entity Recognition Evaluation (2)

Our result is the ninth place, and our nickname is *green*:

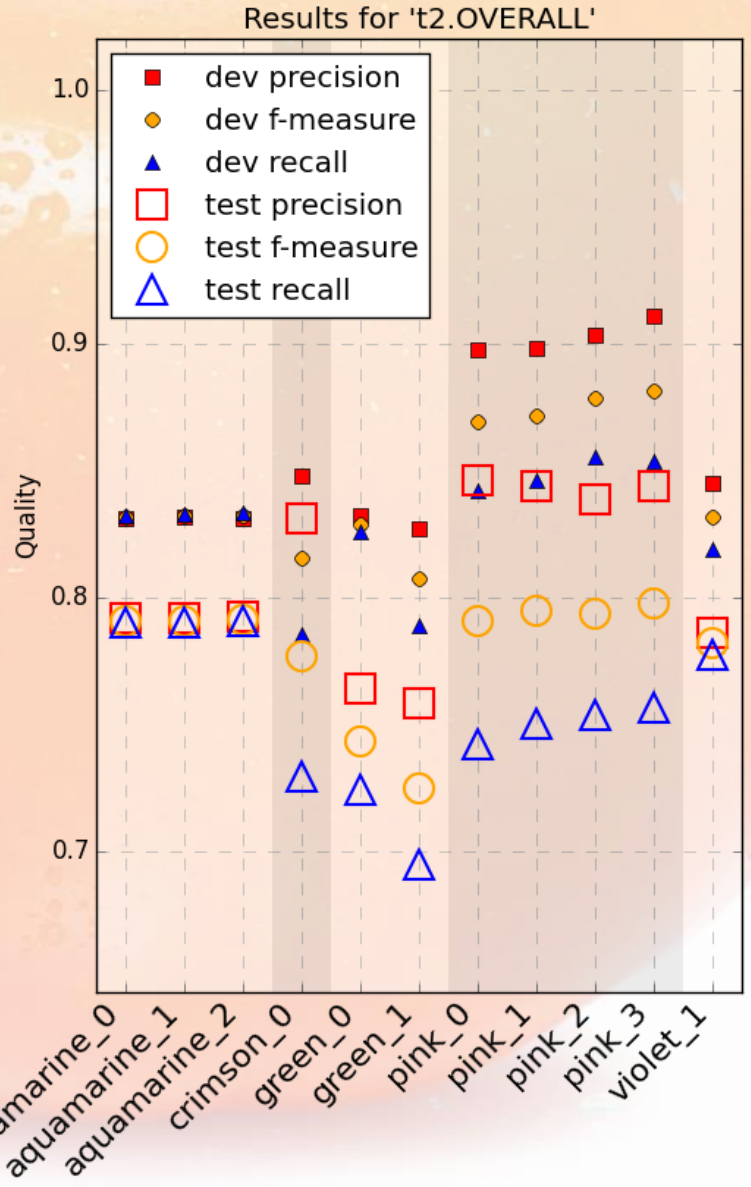
Entity type	**Our Precision**	**Our Recall**	**Our F-measure**	**Best F-measure**
Persons	0.9300	0.8403	0.8829	0.9300 (violet)
Locations	0.9535	0.8361	0.8910	0.9087 (beige)
Organizations	0.8181	0.5450	0.6542	0.7858 (violet)
OVERALL	0.9038	0.7301	0.8077	0.8672 (violet)

Entity Normalization Evaluation (1)

- Best overall F-measure:

  0.7979 (pink);
- Our overall F-measure:

  0.7439 (green).



Entity Normalization Evaluation (2)

We decided to try to evaluate our normalization techniques not taking into account our recognition quality and yielded a satisfactory result:

Entity type	Our Precision	Our Recall	Our F-measure	Rel. F-measure
Persons	0.8024	0.8433	0.8223	0.9313
Locations	0.9017	0.7741	0.8330	0.9349
Organizations	0.6490	0.5760	0.6103	0.8359
OVERALL	0.7725	0.7173	0.7439	0.9210

Fact Extraction Evaluation (1)

Type	Subtype	Test set		Development set	
		Count	%	Count	%
Named Entity	Persons	1347	32	728	31
	Organizations	1537	37	661	28
	Locations	1283	31	943	41
Fact	Ownership	141	23	17	7
	Occupation	336	54	180	78
	Meeting	45	7	5	2
	Deal	102	16	29	13

Fact Extraction Evaluation (2)

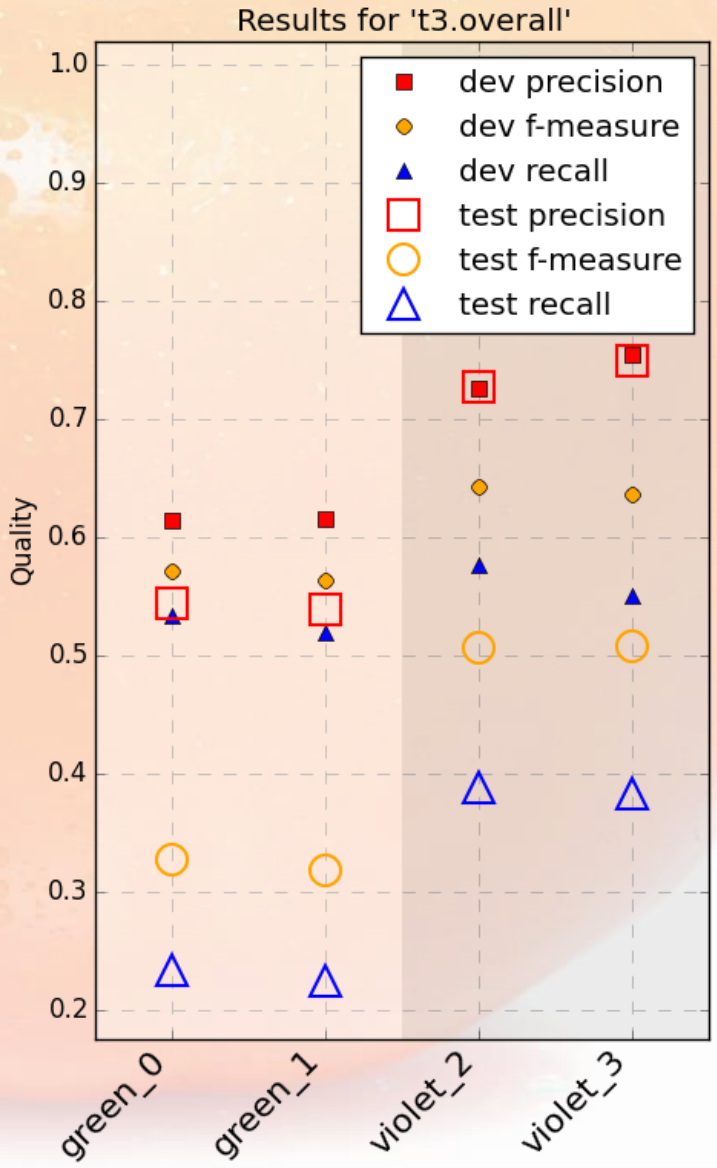
- The Issue here: test and development sets are highly unbalanced;
- Dominant fact type in development set is Occupation, which can be extracted with simple context rules similar to our NER extraction rules.

Fact Extraction Evaluation (3)

- Best overall F-measure:

  0.5078 (violet);
- Our overall F-measure:

  0.3273 (green).



Thank you!
Any questions?

Please, visit our web site at

hurmining.com