Dictionary-based Ambiguity Resolution in Russian Named Entities Recognition. A Case Study

СЛОВАРНЫЙ ПОДХОД К РАЗРЕШЕНИЮ ОМОНИМИИ ПРИ ВЫДЕЛЕНИИ ИМЕНОВАННЫХ СУЩНОСТЕЙ В РУССКОМ ЯЗЫКЕ

Maria Brykina m.brykina@gmail.com
Alexandra Faynveyts fainalex@yandex.ru
Svetlana Toldova toldova@yandex.ru

Outline

- NE Recogniton task: task specification
- Problem: NE ambiguity types; possible solutions
- System architecture and functionality
- Dictionary based module
- Basic homonymy NE cases and dictionary filling rules
- Results and Conclusion

Named Entities Recognition and Identification

Eventos

Objects

All

Person

- Обама Барак
- Патрушев Николай ✓
- Путин Владимир
- Шойгу Сергей
- Джордж Литтл
- Кэйтлин Хэйден
- Томас Донилон
- Юрий Ушаков

Географическое ...

- Вашингтон
- Ирландия
- Москва
- Оклахома
- Россия

Properties

Патрушев Николай	
imageUrl	
givenName	Николай
additionalName	Платонович
lastName	Патрушев
статья Википедии	http://ru.wikipedia.org/wiki/Николай_Патрушев
birthDate	11.07.1951
birthPlace	Ленинград
EmployedBy	Совет Безопасности РФ
EmployedBy	ФСБ РФ
title	Герой Российской Федерации
title	генерал армии
title	доктор юридических наук

Text Политика

Развитие отношений между США и Россией, в том числе в экономической области, обсудили президент Барак Обама и секретарь Совета безопасности РФ Николай Патрушев. Встреча состоялась 22 мая в Белом доме и на ней, помимо прочего, затронули вопросы борьбы с терроризмом и ситуацию в Сирии.Как сообщила официальный представитель Совета национальной безопасности США Кэйтлин Хэйден, Обама заглянул на встречу Патрушева с помощником президента США по национальной безопасности Томасом Донилоном."Президент США подтвердил желание укреплять двусторонние отношения, в том числе американо-российские экономические связи. Они также говорили о важности углубления сотрудничества в борьбе с терроризмом и необходимости политического урегулирования в Сирии путем переговоров",

Task Specification

- Input: Russian news texts
- Restricted (user-specified) target set of NE
- High precision required
- The system should be easily adapted to new sets of NE
- Any updates should result in the same precision of “old” objects

Possible domains of application: editorial interfaces in news agencies, tagging of large archives with NE

Problem: NE Ambiguity

- NE vs. common noun:
  Rubin (FC)
  Rubin (jewel);
- NE class 1 vs. NE class 2:
  Vladimir (town)
  Vladimir (personal name);
- ontology ambiguity between two entities with the same name:
  Sergei Ivanov (politician)
  Sergei Ivanov (scientist)
  … (18 more in Wikipedia)

Possible Solutions

- Machine learning-based systems
- **Handmade rule-based systems**
  - predictable system behavior
  - high precision for the user-specified list of NEs
  - the possibility for the non-specialist to update the system
  - the possibility to control the effects of NEs database extension

System Architecture and functionality

Eventos

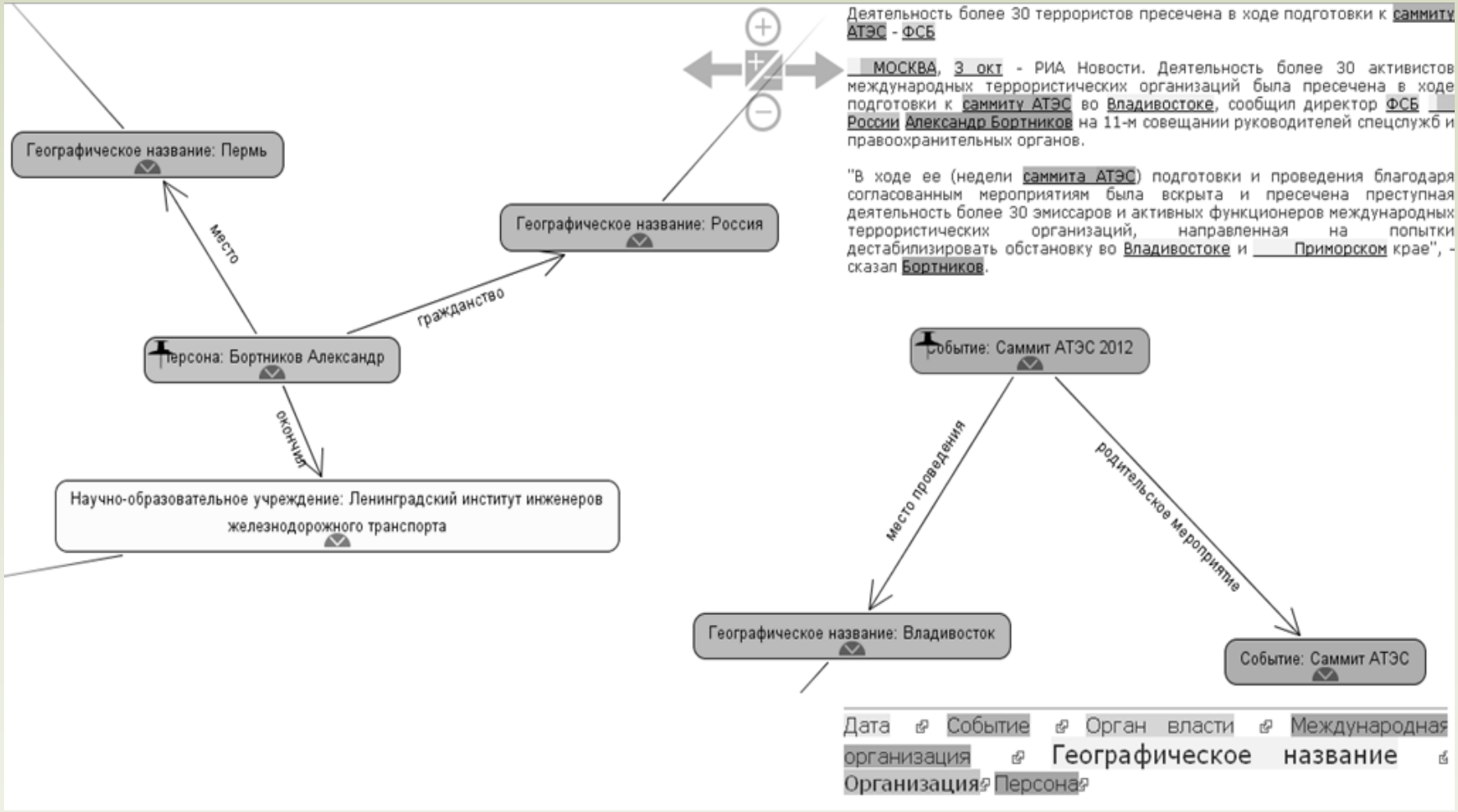
OntosMiner processor

- Domain ontology
- Basic resources (tokenization; lemmatization; sentence splitter; numbers, dates, currency extraction)
- **Dictionary-based NE extraction**
- Heuristic and statistical PLO extraction
- Positional NE class ambiguity resolution

JAVA; GATE; triples of a specific format Turtle; RDF-storage

OntosMiner Output Visualization

Eventos



Dictionaries Structure (1)

- Objects (NE) dictionaries: objects (ontological information)
- Synonyms dictionaries: lists of synonyms with reference to an object ID

Main principles

- we want to avoid constructing huge full lists of all synonymous noun phrases

  Vladimir Putin *Vladimir Vladimirovič Putin*

  V. Putin *V.V.Putin etc.*

- we use a minimum common text material for a set of synonyms

  Putin

Dictionaries Structure (2)

Eventos

Synonyms:

ambiguous vs. unambiguous

need to be immediately verified through the context

Tri bogatyrja

identify an object

UNESCO

Dictionaries Structure (3)

Verification

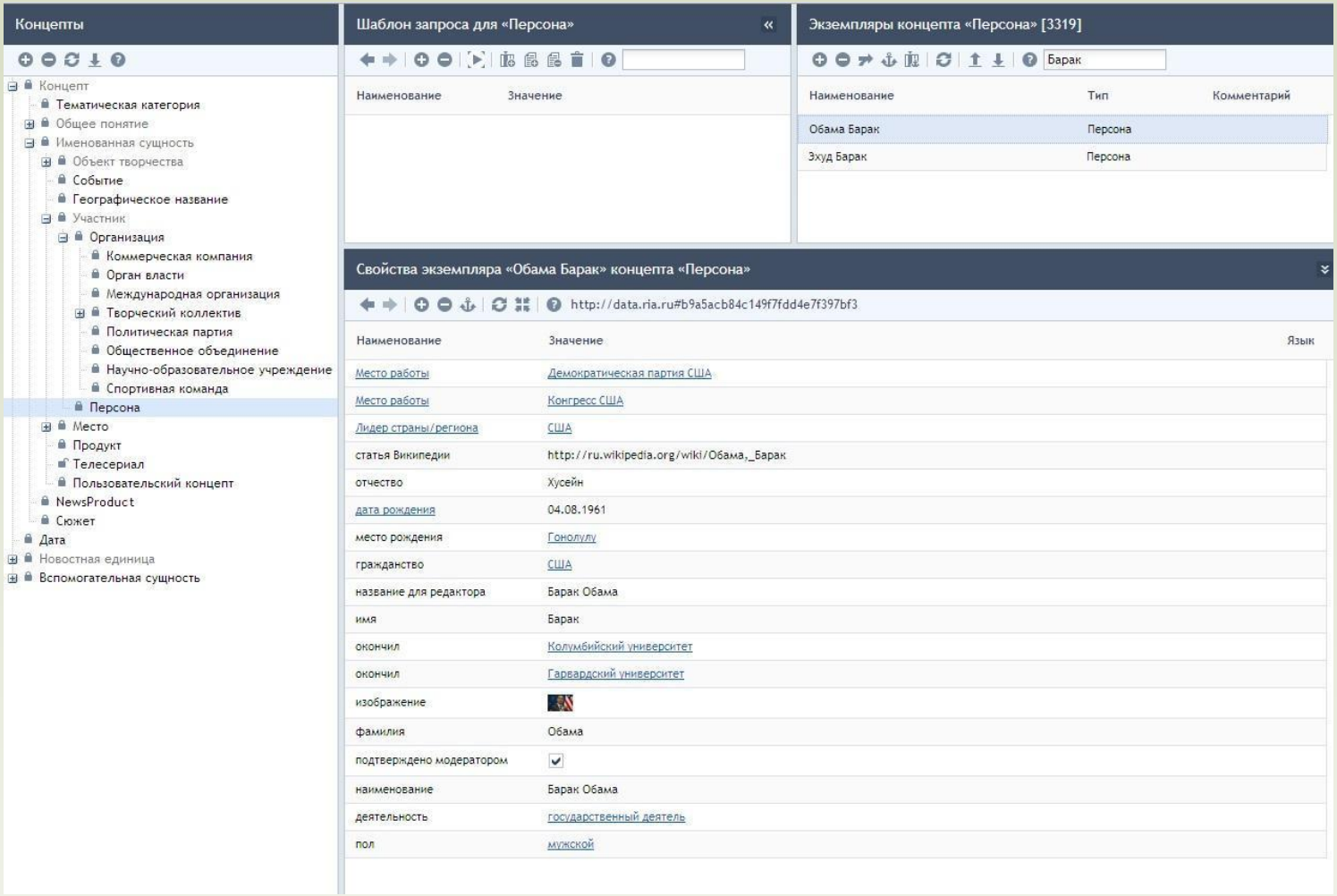
- Verification of ambiguous synonyms is performed by manually assigning attributes to them that specify the additional information needed to resolve ambiguity.
- These attributes are then processed with rules.

Questions

- What kind of information is needed for various classes of NE?
- Is it possible to have a unified template for an arbitrary objects (NE) class?
- How does one evaluate such a system?

Dictionaries Structure (4)

Eventos



Organizations (1)

Eventos

i. Organization vs. other entity ambiguity

- punctuation

OR

- prefix (left-adjoined word)

Lookup entry	attribute	entity
Mir dereva Mir dereva (lit.: the world of wood)	*needs quotation marks;* *needs left-adjoined word (sequence) from the list*	*Derevoobrabatyvajushchaja kompanija «Mir dereva»* Woodworking company "Mir dereva"

Organizations (2)
prefix OR punctuation

Eventos

- OK *«Mir dereva» javljaetsja krupnym proizvoditelem izdelij iz dereva.*

  “Mir dereva” is a big manufacture for wooden products.
- OK *Kompanija Mir dereva - odin iz organizatorov ètoj vystavki. -* -

  “Mir dereva” company is one of the organizers of this exhibition.
- *Tema segodnjashnego zanjatija v detskom sadu «Mir dereva i metalla»* -

  Theme of the today’s lesson in the kindergarten is “The world of wood and metal”

Organizations (3) identical names problem

ii. Organization vs. Organization ambiguity-1

Rubin (FC) vs. Rubin (design office)

- punctuation doesn't help
- common words or phrases like *organization* or *CEO* don't help
- we need to specify an industry to which a corresponding NE refer

Organizations (4)
Location problem

Eventos

iii. Organization vs. Organization ambiguity-2

Organizations with identical names can be situated in different regions:

Ministry of Foreign Affairs of Russian Federation

Ministry of Foreign Affairs of France

We need to take into account the nearest mention of location in the document

If no location is found in text or metatext information default locality - Russian Federation is set to some government structures.

Organizations (5)

Eventos

- Ministerstvo inostrannyh del RF otvetilo na pros'bu Onishchenko
  Ministry of Foreign Affairs of Russian Federation answered to the request of Onishchenko.
- *Ministerstvo inostrannyh del otvetilo na pros'bu Onishchenko*
  Ministry of Foreign Affairs answered to the request of Onishchenko.
- *Vladimir Putin posetil Parizh i provel vstrechu s glavoj MIDa*.
  Vladimir Putin visited Paris and hold a meeting with the head of the Ministry of Foreign Affairs.

Organizations (6)
Controlled Vocabularies

Eventos

- ***prefix/postfic*** designates the following AL as an organization

general prefix	*specific prefix*
doesn't imply a specific industry	implies a specific industry
company, society	*bank*

- ***keyword*** means that the AL is *probably* an organization (*director, CEO*)
- ***key adjective*** forms a specific prefix when combined with a general prefix (*cosmic, aircraft*)
- ***postfix*** designates the previous AL as an organization (*Limited, & Co*)

Locations (1)

i. Location vs. Location ambiguity

8 Soviet regions in Russia

- ontological information about parent locations or locations of the same level is used to calculate an appropriate variant from the text

Tunis vs. Tunisia

- system of prefixes (city, country, ...)

ii. Location vs. Common words ambiguity

g. Nahodka – lit. (the city of) Find vs. nahodka

- prefixes or close locations
- PP groups

Locations (2)

Eventos

iii. Location vs. Person ambiguity

Mogilev – Mogilev City

Lion Izmajlov vs. Lion (city)

- heuristic module which extracts Person entities including those not present in a given ontology

  +

  minimization rule: when a Location is embedded into a Person, this Location is removed

Events (1): what is it?

- exhibitions
- film festivals
  - *Festival de Cannes*
- messages of the President
  - *messages of the President to the Federal Assembly*
- summits
  - *United Nations summit*

Events (2): left context

i. Event vs. another entity or common word ambiguity

- quotation marks are not used regularly, so they don't help
- left-adjoined words (prefixes) help

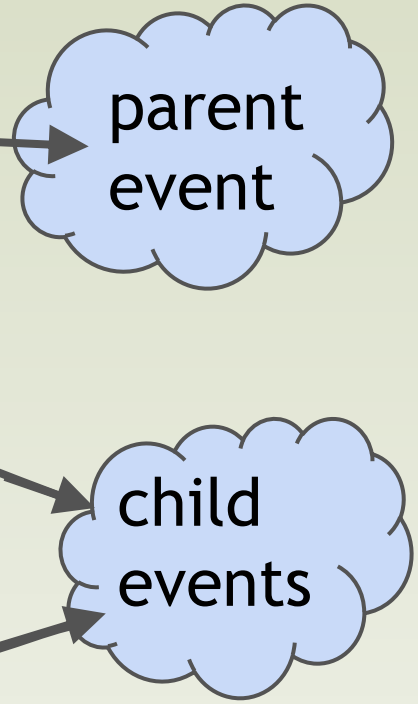
Synonym	attribute	entity
Zolotoj lev Golden lion	needs left-adjoined word (sequence) from the list for the theater festival	Theater festival “Zolotoj Lev” (lit.: Golden lion)

Events (3): parents and childs

Eventos

ii. Parent Event vs. Child Event

- *Evrovidenie – vazhnoe sobytie v mire pop-muzyki.*
  Eurovision is an important event in pop music world.
- *Poslednee Evrovidenie proshlo v Moskve.*
  The last Eurovision took place in Moscow.
- *Na Evrovidenii v 2009 godu pobedil Dima Bilan.*
  Dima Bilan won Eurovision contest in 2009.



Events (4): parents and childs

A periodical Event must be associated with a unique **place**, **number** or **year** to be recognized as a specific child Event

synonym	attributes	entity
Evrovidenie Eurovision	needs a time-marker: year (2012) OR needs a place-marker (Baku)	*Evrovidenie-2012* Eurovision-2012

Events (5): AND

Sometimes we need to use AND-operator between "needs a time marker" and "needs a place marker" attributes.

Lookup	attributes	entity
avtosalon automobile show	needs a time-marker: year (2011) AND needs a place-marker (Toronto)	*Avtosalon v Toronto 2011* Automobile show in Toronto – 2011
avtosalon automobile show	needs a time-marker: year (2012) AND needs a place-marker (Toronto)	*Avtosalon v Toronto 2012* Automobile show in Toronto – 2012

Events (6): AND

Eventos

Avtosalon v Toronto vsegda ochen' interesnyj.

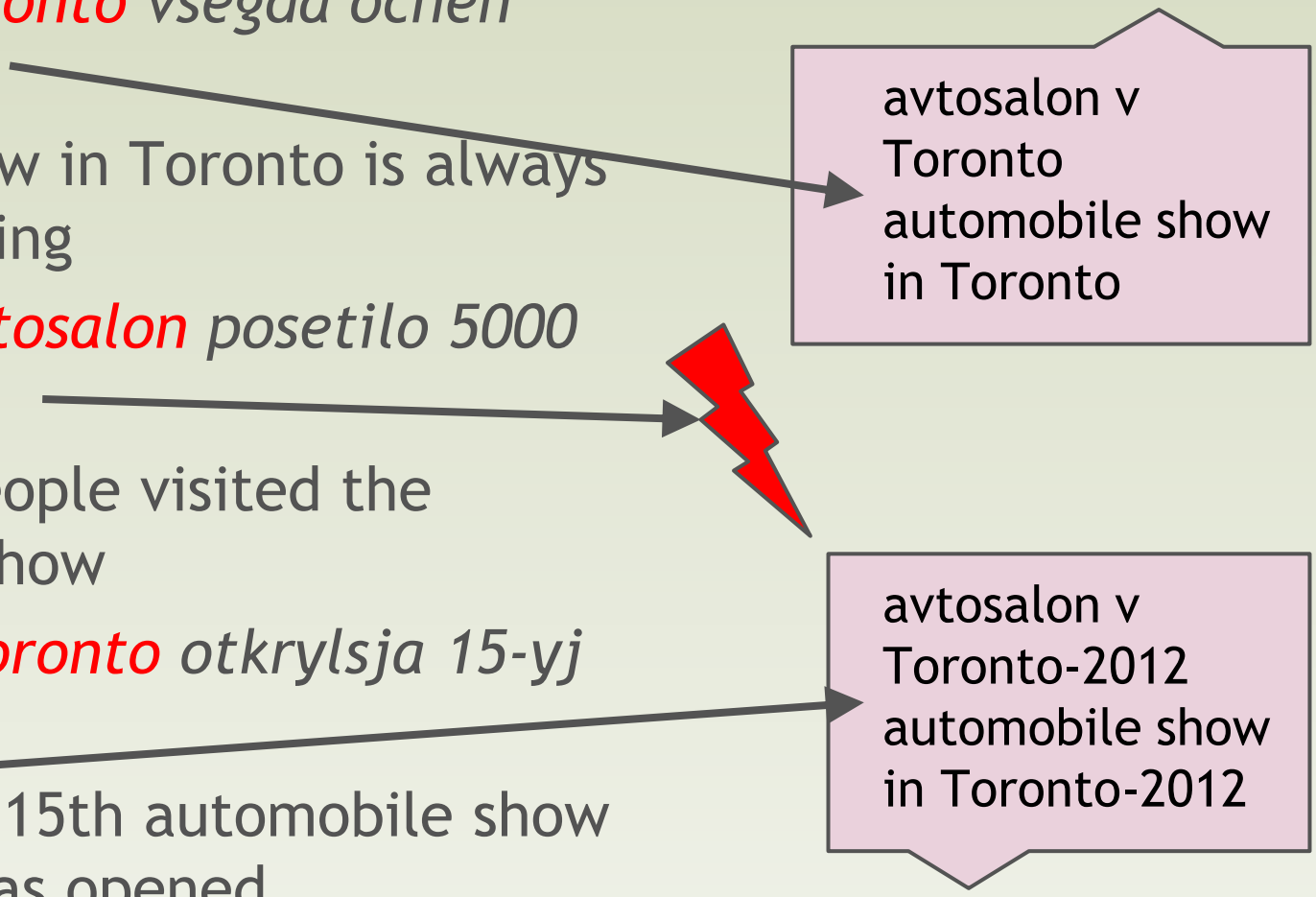
Automobile show in Toronto is always very interesting

V 2012 godu avtosalon posetilo 5000 chelovek.

In 2012 5000 people visited the automobile show

03.04.2012 v Toronto otkrylsja 15-yj avtosalon.

03.04.2012 the 15th automobile show in Toronto was opened



Unified Attributes Template

Q: Is it possible to have a unified template for an arbitrary objects (NE) class?

- A set of context attributes

attribute	**scope (to choose)**
needs an object	Left or right context/ Sentence/ Paragraph/ Document
needs a substring	
stop object	
stop substring	

Unified Attributes Template (2)

Eventos

lookup	attribute	entity
sbornaja Rossii national team of Russia	needs this object	*sbornaja Rossii po hokkeju* national hockey team of Russia
sbornaja Rossii national team of Russia	needs this object	*sbornaja Rossii po futbolu* national football team of Russia

Unified Attributes Template (3)

Attributes can be combined with each other with AND and OR operators

- Users can set whether quotation marks can be used or are required to identify an object
- Users can add ALs that are verified only if a corresponding entity has already been found in the text

Evaluation (1)

Eventos

NE	*Number of cases*	*Recall*	*Precision*	*F-measure*
Location	2270	0,98	0,99	0,98
Organization	1654	0,93	0,95	0,94
Person	453	0,94	0,99	0,96
Event	55	0,98	0,85	0,91
Product	138	0,96	0,98	0,97

Evaluation (2)

- we only provide counts for NPs that correspond to the NE from the Domain;
- the quality of the system depends crucially on the quality of manually filled dictionaries, and thus these results cannot be reproduced unless based on the same dictionaries;

However:

- the lightweight system based on user dictionaries and rather simple rules could be quite helpful if one's goal is extracting a limited number of NEs;
- the cases of ontology ambiguity are not too frequent in general news texts to influence significantly the performance of the system

Conclusion (1)

Clues for NE disambiguation:

- text properties (quotation marks);
- ontological properties:
  - object class hierarchy;
  - location associated with an object;
  - class-specific features (status-role of a Person, organization industry etc.).

Clues may have different scope, e.g. a sentence, a paragraph, the whole text…

Conclusion (2)

In dictionary- and ruled-based systems for NE identification in Russian news texts:

- it is possible to suggest a unified template allowing to add attributes that serve to disambiguate different types of objects (other than PLO);
- along with dictionary-based extraction, it can be useful to have a guessing-module to help resolve some types of ambiguity;
- there can be specific rules based on metatextual information, which recover implicit attributes (for example, Russia can be set as default location for Russian news agencies texts).

TRY IT

Eventos

http://dix.ontos.ru/dix/bookmarklet.jsp

Eventos

Thank you!