PRACTICAL ASPECTS OF LONG-TERM ONTOLOGY-BASED INFORMATION EXTRACTION

Образец подзаголовка

Anna Kravchenko, Vasiliy Pivovarov, Alexander Zharikov

OBIE systems

‘Ontology-based information extraction’ is a subfield of information extraction, where ontologies play an essential role in the process, shaping both system input and target output.

Examples: iDocument, OntoSyphon, ontoX, PoolParty Extractor, SCOOBIE, smart FIX

Образец подзаголовка

SCAN

Scan.interfax.ru is an entity-oriented news analysis system.

We found that, while there are a lot of articles concerning an OBIE system architecture, we encountered difficulties that were not mentioned anywhere. Architecture that shows good performance in a single test does not necessarily perform as good in the long term.

Образец подзаголовка

Discovered issues

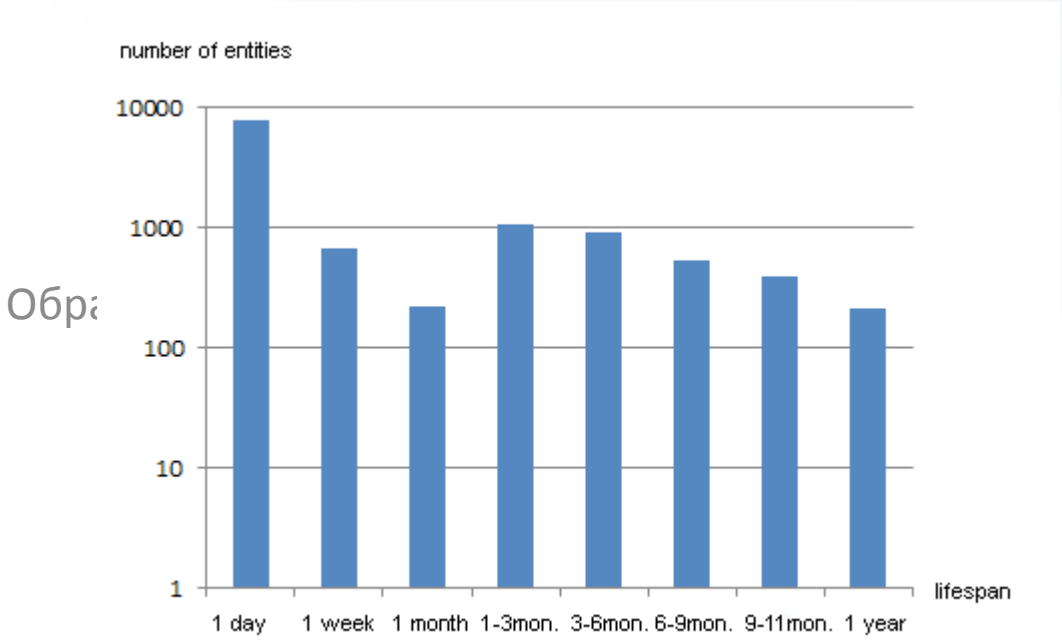
World dynamics is more important than it is often considered.

- New names appear in the news all the time. Some roles (e.g. «British Prime Minister») shift from one person to another
- It is necessary to update the database timely.
- Editing database manually requires a lot of human effort. Automatic updating seems to be the solution.

Образец подзаголовка

Discovered issues

Life span of most entities doesn’t exceed one day.



Discovered issues

However, automatic updating leads to two other problems:
1. Long tails
2. Error accumulation

Образец подзаголовка

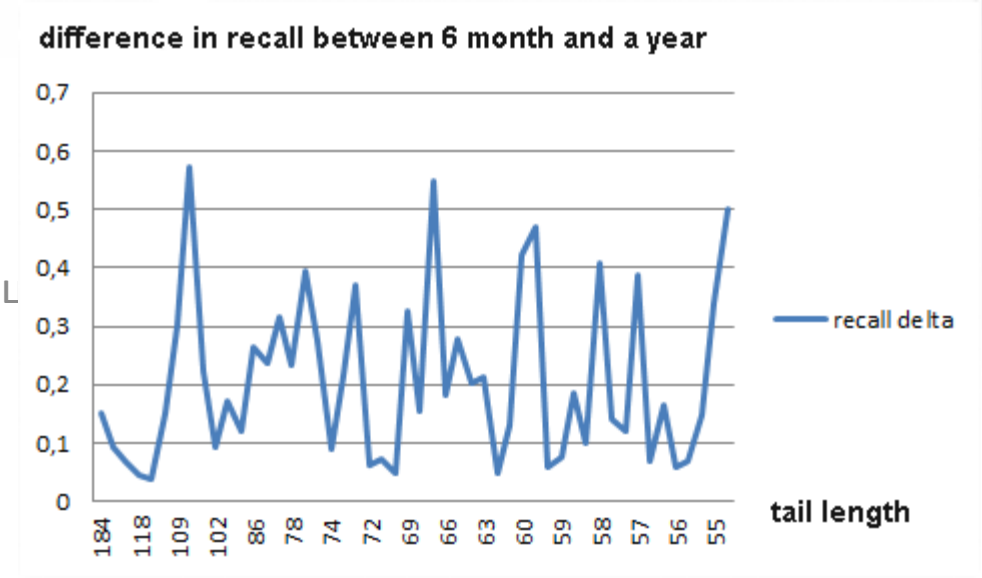
Long tails

The entity extraction procedure can fail to connect person name with a role or to extract a full name. Persons with the same name and different roles are also considered different entities and even simple roles are difficult to merge.

- Muammar Gaddafi: (Muammar, Mummar, Muamar…)x(Caddafy, Kaddafy, …).
- Jim Jarmush:“acclaimed director and musician” and “the creator of the film “Limits of control””. Образец подзаголовка

Long tails

It strongly influences the recall rate and can potentially slow down the system.



Образец

6.6.14

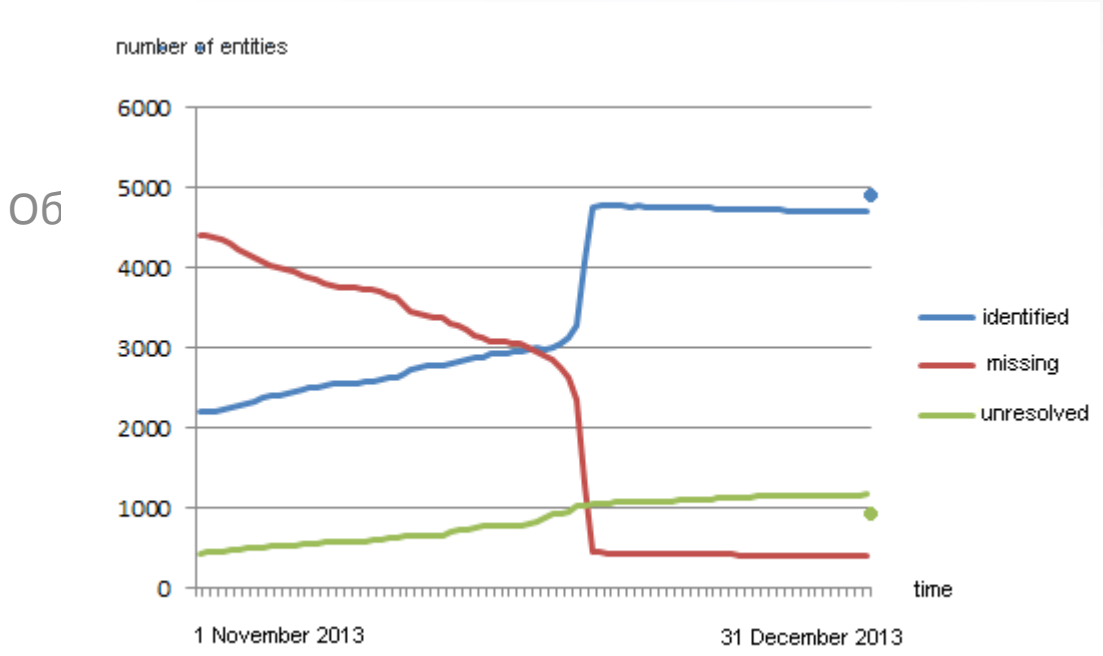
Error accumulation

- A procedure with 99% of precise identifications creates an impression of almost perfect.
- However, in the long term it leads to a burst of identification errors, because ontology is being filled with incorrect entities.
- For automatically updated ontologies we consider this problem the most serious.

Образец подзаголовка

Proposed solutions

To evaluate the degradation rate for the object identification system we counted change of the number of precise entity identifications, the number of unresolved ambiguities and the number of identifications missing (candidates for identification were not found in the database) for a given time interval. The speed of the identification degradation due to ambiguity growth (the slope angle of the green line on the graph) seems to be a good measure of quality, which can be used to test overall system quality and internal consistency and perhaps to compare different systems between each other.



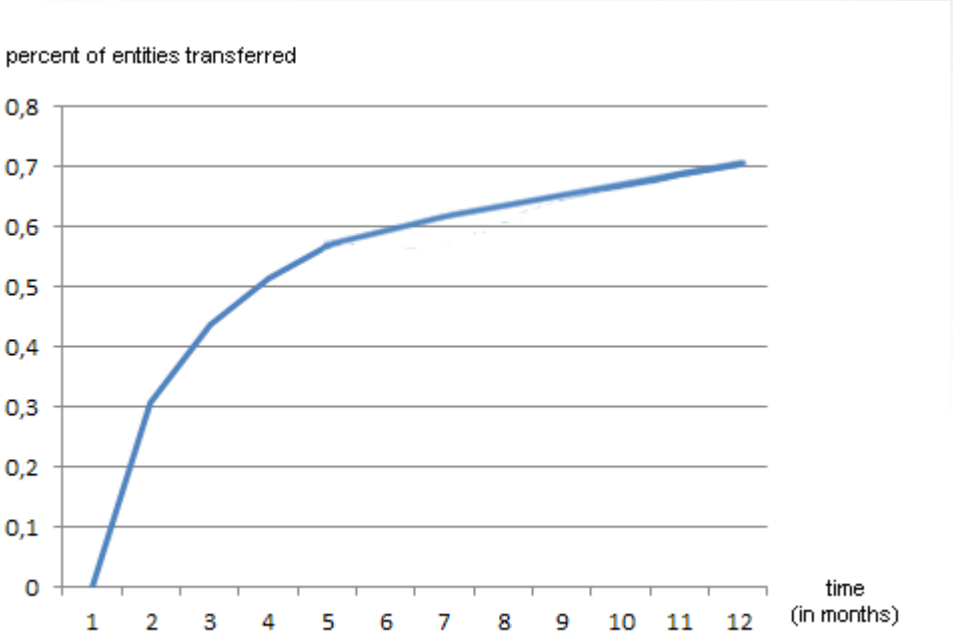
Proposed solutions

- We apply a merging algorithm to our database. For example, we consider people with similar names belonging to the same organization the same person.
However, it is important that organizations can't be merged automatically, only persons.
- To deal with the accumulating errors we propose using separate databases for each time period (for example, a year).
Important entities are repeated often while mistakes are rarer and filling a new database allows to "clean up"

Образец подзаголовка

Proposed solutions

Percent of entities transferred to the new database during the year:



Образе

Results

This logic was successfully implemented in Scan.interfax.ru project and has provided sufficient quality.

Companies:
- 99% precision and 78% recall for raw entities
- 100% precision and 95% recall for “active” entities, verified by human

Persons: Образец подзаголовка
- 90% precision and 90% recall for raw entities
- 95% precision and 97% recall for “active” entities

Questions?

Образец подзаголовка