# PRACTICAL ASPECTS OF LONG-TERM ONTOLOGY-BASED INFORMATION EXTRACTION

**Kravchenko A.** (anna.kravchenko@interfax.ru),
**Pivovarov V.** (vasiliy.pivovarov@interfax.ru),
**Zharikov A.** (alexander.zharikov@interfax.ru)

Interfax, Moscow, Russia

'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. There are many different approaches to creating and maintaining an ontology and little work has been done to evaluate and compare the effectiveness of those approaches.

In addition, the practical applications of those systems differ drastically from theory. Architecture that shows good performance in a single test does not necessarily perform as well in the long term.

We conducted an experiment to explore the issues that arise during practical application of OBIE methods and to describe the behavior of ontologies maintained during a long period of time.

In this article we discuss emerging problems and propose working solutions for them as well as the way of evaluation of OBIE systems. Those solutions were successfully implemented in the scan-interfax.ru project and have provided sufficient quality for the commercial use of an advanced entity-based search engine extracting information from news.

**Keywords:** ontology, information retrieval, search engine, ontology-based information extraction, news extraction, news analysis, ontology population

## 1. Introduction

The term 'ontology-based information extraction (OBIE) only appeared a few years ago, though some work related to this field has been carried out much earlier. It is a subfield of information extraction, where ontologies play an essential role in the process, shaping both system input and target output.

Information extraction (IE) mostly deals with shallow parsing of the processed data, without attempting a deep linguistic analysis of all aspects of a text. In this way IE systems can be sufficiently fast to deal with the large amounts of web data. At the same time, the text itself may contain conceptual structures and semantic links that are crucial for understanding its meaning and need to be processed thoroughly, especially for domain-specific tasks. Using ontologies allows to combine both those approaches.

Evidently, the quality of the ontology used is critically important for such system to work. There are different ways to create an ontology, a good overview is given in [Wimalasuriya, 2010]. In most cases ontologies are created manually or taken off-the-shelf, some use automatically populated ontologies. Both options have their benefits. For example, a manually created ontology is better for identifying geographical names, while news articles require constant influx/addition of new data due to world dynamics.

It also should be noted that practical applications of those systems differ drastically from theory. Architecture that shows good performance in a single test does not necessarily perform in a satisfactory manner in the long term. New entities that are absent in fixed/manually created ontologies appear constantly in the world, and for automatically updated ontologies errors tend to accumulate.

We've conducted an experiment to explore the issues that arise during practical application of OBIE methods and to describe the behavior of an ontology maintained during a long period of time.

In this article we discuss issues that emerged during the experiment and propose working solutions for them, as well as a way of evaluating OBIE systems.

## 2. Experiment details and system architecture

Most rule-based systems (except for [Hwang, 1999]) use manually constructed ontologies which are not updated. Updating ontology automatically increases recall and also provides opportunity for research.

For our experiment we used the scan.interfax.ru system, which is focused on news analysis and entity-oriented search. The demo version is free, though available with limited functionality.

The system is mostly rule-based to maintain precision, although it uses some statistical algorithms such as Bayesian and SVM classifiers. Ontology classes and relations (Tbox) are set up manually while the collection of entities (Abox) is constructed and updated automatically from news articles texts using a bootstrapping approach. Scheme of the system's architecture can be seen on Fig. 1.

The procedure of adding entities to the database consists of two stages.

In the first stage entities are extracted from the article. It is a part of a more general procedure, which also reveals morphological and syntactic structures, conducts anaphora resolution, extracts keywords and key sentences, etc. The key feature of this entity extraction approach is its independence from any time-dependent object databases. In other words, the system aiming to extract person or organization entities does not use any lists of real world's ones and only operates with semantic and morphological dictionary data.

At second stage extracted essences are identified with an entity stored in the database. If such an entity is not found, a new one is created. The stored entity should represent a real world object and has links to any information about it.

A detailed description of the algorithm can be found in [Zharikov, 2011].

For our experiment we chose a time interval from 01.01.2013 to 31.12.2013 containing approximately 2,500,000 documents from the news stream. The identification

of entities was held with only rule-based precise algorithms used (all statistical procedures were switched off). At its initial state the database was empty. At the end of the procedure there were approximately 1.2 million of entities (persons and organizations) collected.

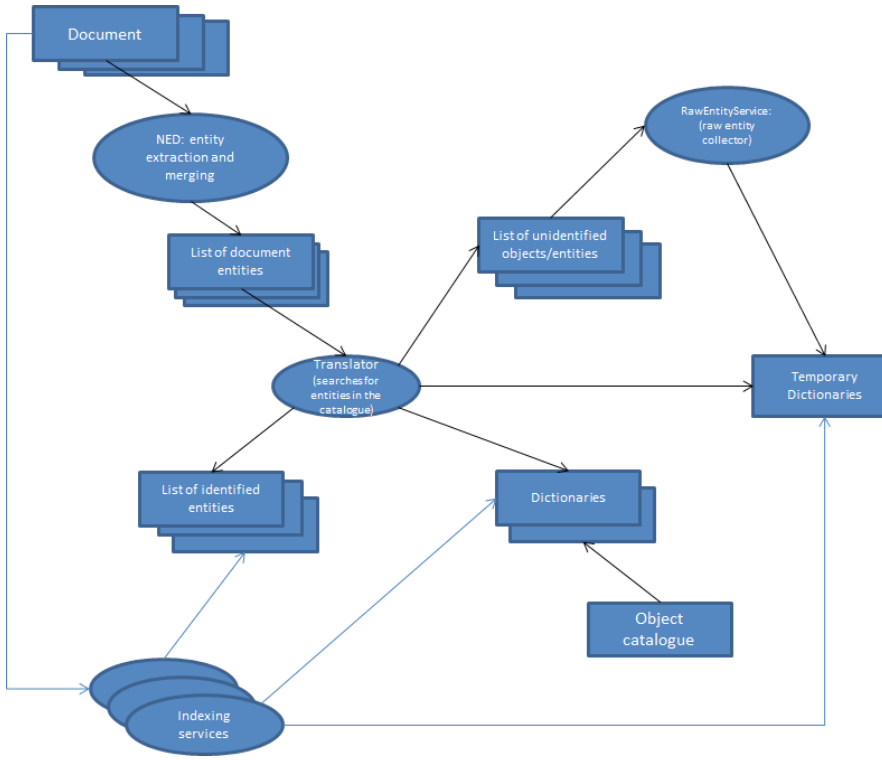


**Fig. 1.** Process of entity extraction in Scan system

Three values were measured: number of precise entity identifications, number of unresolved ambiguities and number of identification missing (candidates for identification were not found in the database).

We also measured the lifespan of extracted entities and their number of occurrences.

## 3. Discovered issues

In this article we will analyze the extraction of person entities since they are more illustrative. In context of analyzed issues organization extraction specifics is very similar.

### 3.1. World dynamics problem

World dynamics is the main source of ambiguity. It is also unavoidable.

New names appear in the news all the time. Also some roles shift from one person to another, for example, “The British Prime Minister” may refer to different people during different time periods. A new person with the same name may appear or the person could change his or her role. All this leads to ambiguity of interpretation.

Ideally, date range for each entity should be stored in the database, and the list of relevant entities should be generated accordingly to the date assigned to the processed document. The database should also be updated timely.

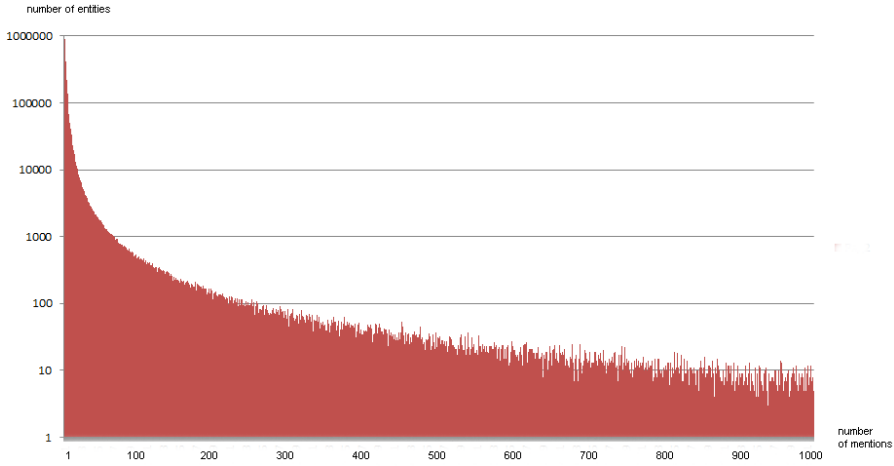


**Fig. 2.** Number of entities corresponding to number of mentions

As we can see on Fig. 2, most of the entities are mentioned only a small number of times or even once. Even further, Fig. 3 shows that the lifespan of most entities doesn’t exceed one day.

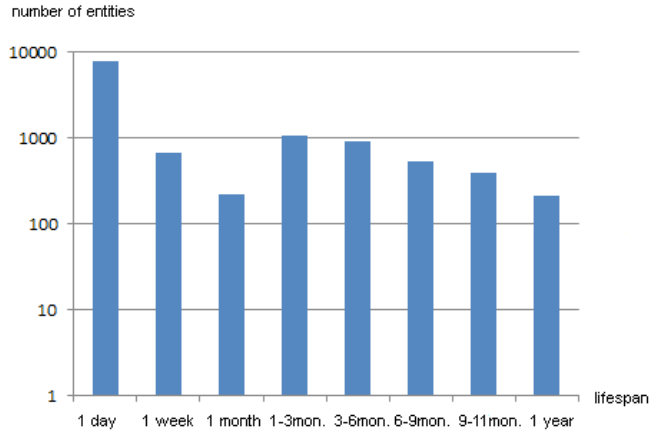


**Fig. 3.** Number of entities corresponding to their lifespan (we removed entities that were mentioned less than 20 times from this chart)

We should also note that every entity belongs to one of the three following types.

1. Rare entities
   Number of mentions is very small (less than 10 mentions usually). A good example would be a school headmaster that sometimes appears in the regional press. Lifespan is fairly long.
   Those entities usually come along with proper qualifiers and are easy to extract correctly.
   Qualifiers (not to be confused with qualifiers in formal semantics) are noun groups that describe person's role. For example, in “the great painter Van Gogh” the phrase “the great painter” would be a qualifier.

2. Constant objects.
   Those fall into one of two subcategories:
   a. Contemporary public figures (mostly politicians), those have high occurrence and long lifespan.
   b. Historical figures, are characterized by low/medium occurrence and very long lifespan, usually exceeding system’s lifespan.

   It is hard to extract roles for those persons, because those roles are considered common knowledge and a mention of such a person usually comes without a proper qualifier.
   However, using an ontology allows to extract such entities successfully, since in a large enough corpus it is eventually possible to find a good qualifier. Nevertheless it is recommended to have a small list/dictionary of historical persons, because some of them may be rather rare.

3. Cluster objects
   Objects of this type have high occurrence and short lifespan. They are usually connected to the same news story.
   They mostly appear with roles and are easy to. To enhance recall it is also recommended to use topic detection methods.
   Cluster object are the main reason to update database timely.

A similar classification can be proposed for organization entities:

1. Rare or common entities such as local stores or motorcar factories. They usually have long lifespan, small number of mentions and the reader is not supposed to know about them, so they are always mentioned with proper qualifiers and are easy to extract.
2. Constant objects that readers are supposed to know about. They tend to have high occurrence, long lifespan and no qualifiers. Oil companies can be a good example of these. This type consists mostly of present-day companies, since in comparison to historical person objects historically-significant organizations are extremely rare.
3. Cluster objects are exactly the same as person cluster objects.

Relying on this classification, we can conclude that updating the database regularly is important. Editing database manually requires a lot of human effort. Automatic updating seems to be the solution, however it causes two other problems.

### 3.2. Long tails

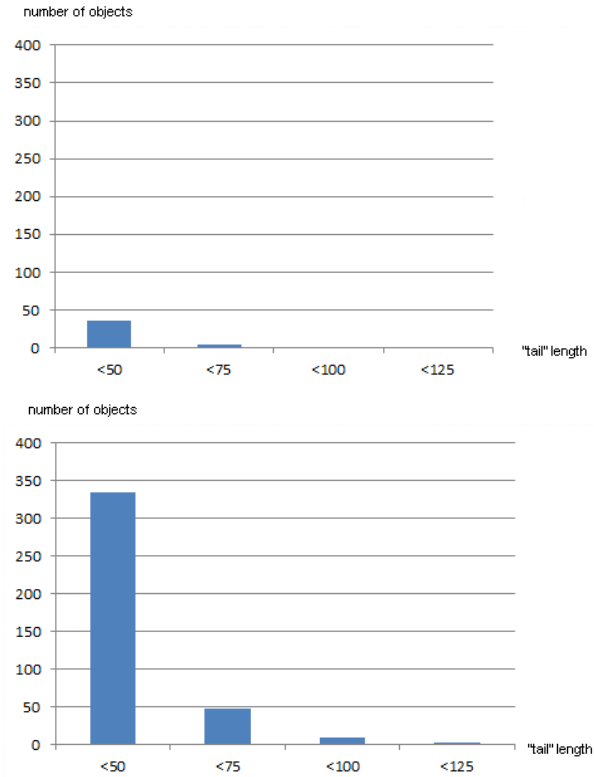


**Fig. 4.** Number of long tails after 3 months and after 6 months of documents

The entity extraction procedure can fail to connect person name with a role or to extract a full name. It can also fail at identifying the entity if it is misnamed. Merging similar names by roles helps, but the same name can correspond to different roles. Persons with the same name and different roles are also considered different entities and even simple roles are difficult to merge, so this creates even more problems.

In automatically updated ontology this leads lead to a long “tail” of entities: for example the tail for Muammar Gaddafi can include a succession of pair derived from (Muammar, Mummar, Muamar…)×(Caddafy, Kaddafy, … ). Tail for Jim Jarmush can contain such roles as “acclaimed director and musician” and “the creator of the film ‘Limits of control’”.

It influences both precision and recall rate (see Fig. 5) and can potentially slow down the system.

This problem can be fixed by a name merging algorithm, as we will show later.

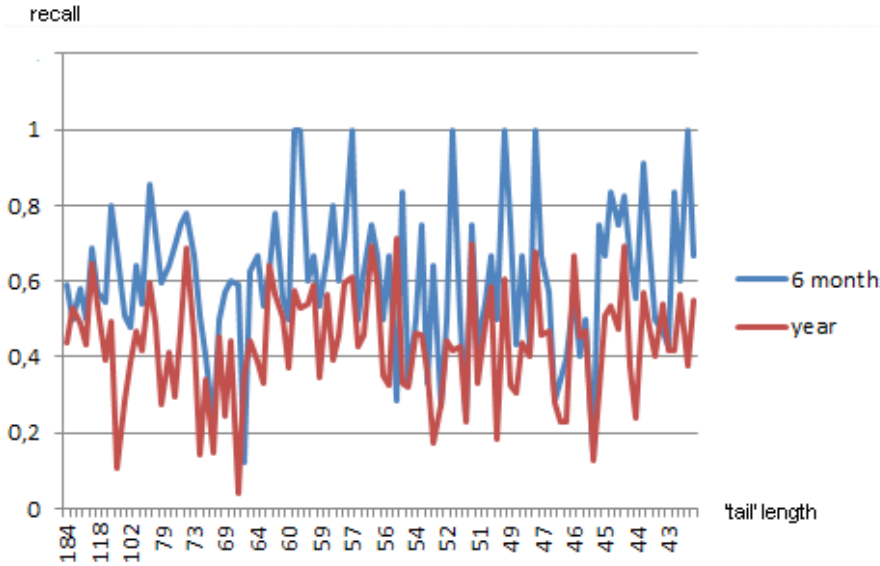


**Fig. 5.** Recall difference for entities with long "tails" after 6 month and a year of database populationg, shown on the x axis is the tail's length. we can see recall values for long-tail entities computed for their mentions in July and for mentions in December, where recall = the number of identified entities divided by the number of matching strings (both found and missing/unresolved entities).

### 3.3. Error accumulation

Existing entity extraction algorithms are not 100% accurate and tend to assign incorrect roles to entities. Even if we obtain an algorithm with 100% precision, the information itself can be unreliable, for example, the author may assign a wrong role to a person. There is also the factor of words having multiple meanings. If you take a geographical dictionary containing several hundred thousands of entries and tag a random text with it, few words wouldn't be considered a geographical object for one reason or another. A similar phenomenon occurs when using a sufficiently large list of organizations. Conflicts arise both within dictionary and with objects of other types.

In practice this inaccuracy becomes crucial. A procedure with 99% of precise identifications creates the impression of being almost perfect, but in the long term it leads to a burst of identification errors, because the ontology is being filled with incorrect entities.

For automatically updated ontologies we consider this problem the most serious.

## 4. Proposed solutions

1. The first and the foremost necessity for a practically applied OBIE system is to update the ontology regularly. The best way is to use an automatically updated ontology. It will help dealing with the influx of new entities in the news stream. However, this may lead to ontology degradation.

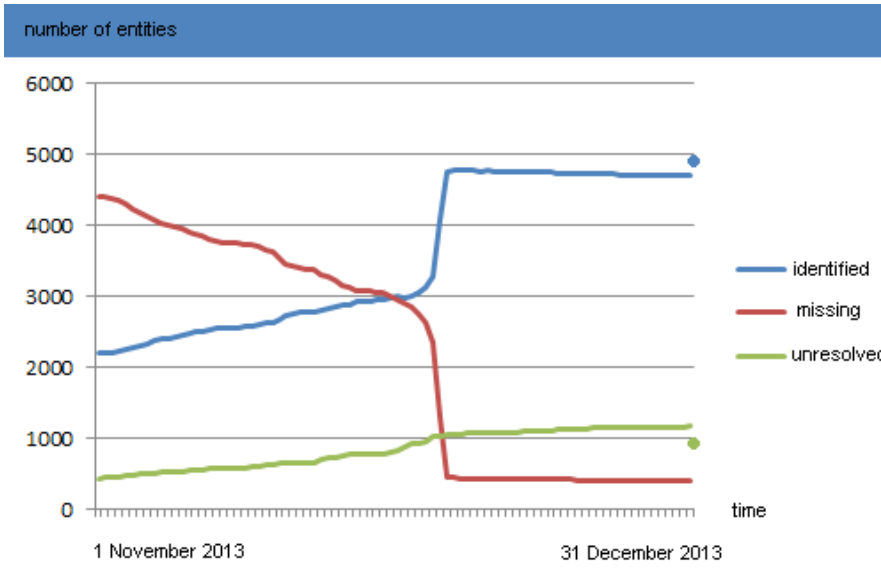


**Fig. 6.** Number of unresolved, missing and successfully identified entities

To evaluate the degradation rate for the object identification system an experiment was performed. Namely, the fixed set of documents was passed through entity identification system under the different states of entity catalog, which corresponds to the increasing values of date. In the certain case considered the document set was a November subset of the general document stream. This circumstance resulted in abrupt grow in the middle of the Fig.6 which corresponds to November period of ontology population procedure

We can see that the number of precise identifications is almost constant and begins to fall very slowly from the middle, while the missing and unresolved rates are inversely related (the growth is the faster, the nearer the database filling procedure approaches the testing interval). We can also see that the growth of unresolved entities rate is nearly linear.

The speed of the identification degradation due to ambiguity growth (the slope angle of the green line on the graph) also 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.

2. Merging of long tails can have significant effect on recall rate.
   The dot on the right end of Fig.6 refers to the number of resolved entities after applying a merging algorithm, based on editing distance and entities' connection to the same organization (people with similar names belonging to the same organization were considered the same person).
3. To deal with the accumulating errors we propose using separate databases for each time period (for example, 6 months).

Important entities are repeated often while mistakes are rarer and filling a new database allows to "clean up" the system from errors and outdated entities.

There is a chance of losing an important entity, but Fig. 6 demonstrates that most frequent entities (the list of entities was chosen by assessor) all have successfully migrated from the older database to a newer one. Entities that appear in the stream from time to time with a sufficiently large interval (months) may still be lost, but this negative effect is not essential for most applications. It can also be compensated for by having a small dictionary of "main" entities, mostly of type 2 (section 3.1) (and by manual database correction ).

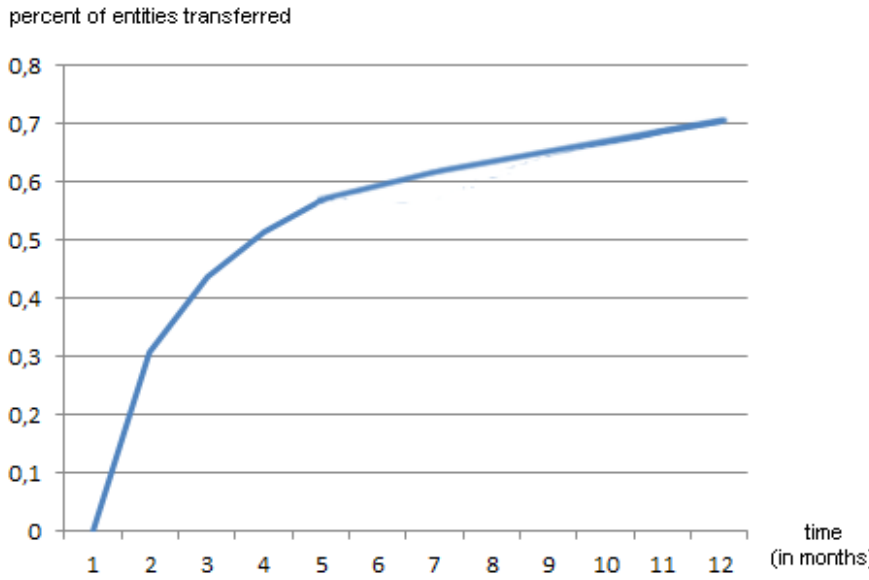


**Fig. 7.** Percent of entities transferred to the new database during the year

## 5. Results and discussion

This logic was successfully implemented in the scan-interfax.ru project and has provided sufficient quality for the commercial use of advanced entity-based search engine.

Performance was analyzed by an assessor. For companies the system achieves 99% precision and 78% recall for raw entities and 100% precision and 95% recall for "active" entities, verified by a human (note, that we do not consider entities splitting in long tails an error), and for persons precision is 90% and 90% recall for raw entities and 95% precision and 97% recall for "active" entities.

## 6. Conclusion

We have described the main aspect of OBIE system behavior in the long term and proposed some approaches to solving emerging problems, which allow us to fill a database with entities from a news stream. Precision provided by algorithms approaches 100% and recall is high enough to set a standart for training a statistical identifier or to make various conclusions automatically. The recall rate is influenced much by the ambiguity caused by world dynamics and by the sensitivity of the system to mistakes in the entity extraction procedure and in the semantic comparison. Nevertheless the negative influence is not crucial for small time intervals (<2 years). The novel approach is proposed to account for these effects by means of using a time dependent database. The negative side of that approach is a loss of the possibility to bring together entities which appear in the stream from time to time with a sufficiently large interval (months). This negative effect is not crucial for most applications. The speed of the identification degradation due to ambiguity growth 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 system between each other. External merge procedures can be implemented periodically to prevent degradation effect

The logic of the precise identification was successfully implemented in the scan-interfax.ru project and has provided sufficient quality for the commercial use of an advanced entity-based search engine. The implementation of supervised methods based on automatically collected standart datasets is in the development stage. The core logic proposed is suitable for a multilingual system though depends much on the entity extraction procedure and entity ontological interpretation which are often language specific. The core logic shown on the example of person entities seems to be suitable for other entity types. The same principles are used in the Scan project to identify organizations and to fill the database with the missing ones (the analog of roles are organization types and locations). The identification of natural language named entities of arbitrary types (such as films, car models, brands at whole, animal names etc.) has been just implemented and now is passing the testing stage.

It includes automatic filling of not only unknown object database but also a base of object types and its semantic hierarchy. The main disadvantages of the proposed approach include lack of universality for different content types and extreme sensitivity to extraction and comparison rules incorrectness.

## References


1. *Zharikov A., Kristalovsky K., Pivovarov V.* Information Retrieval System for News Articles in Russian. Web of Data: The joint RuSSIR/EDBT 2011 Summer School, pp. 5–14 (2011).
2. *Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.* Entity disambiguation for Knowledge Base Population. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 277–285. Beijing, China (2010)
3. *Anastácio, I., Martins, B., Calado P.* Supervised Learning for Linking Named Entities to Wikipedia Pages. In: Proceedings of the Text Analysis Conference, Nov. 2011
4. *Bunescu, R. and Pasca, M.* Using encyclopedic knowledge for named entity disambiguation. In Proceedings of the European Conference of the Association for Computational Linguistic, EACL '06. (2006)
5. *Artiles J., Borthwick A., Gonzalo J.* WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In Third Web People Search Evaluation Forum (WePS-3), CLEF 2010.
6. *Hien T. Nguyen, Tru H. Cao.* Named Entity Disambiguation: A Hybrid Statistical and Rule-Based Incremental Approach. The Semantic Web. Lecture Notes in Computer Sci-ence Volume 5367, 2008, pp. 420–433
7. *Davis A., Veloso A., Soares da Silva A.* et al..: Named Entity Disambiguation in Streaming Data. ACL The Association for Computer Linguistics, pp. 815–824 (2012).
8. *Hassell J., Aleman-Meza B., Budak Arpinar I.* Ontology-driven automatic entity disambiguation in unstructured text. ISWC'06 Proceedings of the 5th international conference on The Semantic Web,
9. *Daya C. Wimalasuriya, Dejing Dou.* Ontology-based information extraction: An introduction and a survey of current approaches. J. Information Science 36(3): 306–323 (2010) [http://www.informatik.uni-trier.de/~ley/db/journals/jis/jis36.html#WimalasuriyaD10]
10. *Chung Hee Hwang.* Incompletely and Imprecisely Speaking: Using Dynamic Ontologies for Representing and Retrieving Information. KRDB, volume 21 of CEUR Workshop Proceedings, page 14–20. CEUR-WS.org, (1999)