

Scenario-Driven Selection and Exploitation of Semantic Data for Optimal Named Entity Disambiguation

Panos Alexopoulos, Carlos Ruiz, and José-Manuel Gómez-Pérez

iSOCO, Avda del Partenon 16-18, 28042, Madrid, Spain,
{palexopoulos,cruiz,jmgomez}@isoco.com

Abstract. The rapidly increasing use of large-scale data on the Web has made named entity disambiguation a key research challenge in Information Extraction (IE) and development of the Semantic Web. In this paper we propose a novel disambiguation framework that utilizes background semantic information, typically in the form of Linked Data, to accurately determine the intended meaning of detected semantic entity references within texts. The novelty of our approach lies in the definition of a structured semi-automatic process that enables the custom selection and use of the semantic data that is optimal for the disambiguation scenario at hand. This process allows our framework to adapt to the particular characteristics of different domains and scenarios and, as experiments show, to be more effective than approaches primarily designed to work in open domain and unconstrained situations.

1 Introduction

Information Extraction (IE) involves the automatic extraction of structured information from texts, such as entities and their relations, in an effort to make the information of these texts more amenable to applications related to Question Answering, Information Access and the Semantic Web. In turn, named entity resolution is an IE subtask that involves detecting mentions of named entities (e.g. people, organizations or locations) within texts and mapping them to their corresponding entities in a given knowledge source. The typical problem in this task is ambiguity, i.e. the situation that arises when a term may refer to multiple different entities. For example, “Tripoli” may refer, among others, to the capital of Libya or to the city of Tripoli in Greece. Deciding which reference is the correct one in a given text is a challenging task which a significant number of approaches have been trying to facilitate for a long time now [2] [3] [6] [7] [5] [8].

The majority of these approaches rely on the strong contextual hypothesis of Miller and Charles [9] according to which terms with similar meanings are often used in similar contexts. The role of these contexts, which practically serve as disambiguation evidence, is typically played by already annotated documents (e.g. wikipedia articles) which are used to train term classifiers. These classifiers

link a term to its correct meaning entity, based on the similarity between the term's textual context and the contexts of its potential entities [8] [10].

An alternative kind of disambiguation evidence that has recently begun to be used are semantic structures like ontologies and Linked Data [7] [6] [12]. The respective approaches typically employ graph-related measures to determine the similarity between the graph formed by the entities found within the ambiguous term's textual context and the graphs formed by each candidate entity's "neighbor" entities in the ontology. The candidate entity with the best matching graph is assumed to be the correct one.

An obvious limitation of this is the need for comprehensive semantic information as input to the system; nevertheless the increasing availability of such information on the Web, typically in the form of Linked Data, can help overcome this problem to a significant degree. Still, however, effectiveness of these approaches is highly dependent on the degree of alignment between the content of the texts to be disambiguated and the semantic data to be used. This means that the ontology's elements (concepts, instances and relations) should cover the domain(s) of the texts to be disambiguated but should not contain other additional elements that i) do not belong to the domain or ii) do belong to it but do not appear in the texts.

To show why this is important, consider an excerpt from a contemporary football match description saying that *"Ronaldo scored two goals for Real Madrid"*. To disambiguate the term "Ronaldo" in this text using an ontology, the only contextual evidence that can be used is the entity "Real Madrid", yet there are two players with that name that are semantically related to it, namely Cristiano Ronaldo (current player) and Ronaldo Luis Nazario de Lima (former player). Thus, if both relations are considered then the term will not be disambiguated. Yet, the fact that the text describes a contemporary football match suggests that, in general, the relation between a team and its former players is not expected to appear in it. Thus, for such texts, it would make sense to ignore this relation in order to achieve more accurate disambiguation.

Unfortunately, current approaches do not facilitate such a fine-grained control over which parts of a given ontology should be used for disambiguation in a given scenario and which not. Some of them allow the constraining of the concepts to which the potential entities may belong [6] [8], but they do not do the same for relations nor do they provide any structured process and guidelines for better execution of this task. That is because their goal is to build scenario and domain independent disambiguation systems where a priori knowledge about what entities and relations are expected to be present in the text is usually unavailable. Indeed, this is the case in scenarios involving news articles, blog posts, tweets and generally texts whose exact content cannot really be predicted. Yet there can be also specialized scenarios where such predictions can be safely made.

One such scenario is the one above about football match descriptions. This was in the context of the project BuscaMedia¹ and involved the disambiguation of football related entities within texts describing highlights of football matches.

¹ <http://www.cenitbuscamedia.es/>

The nature of these texts made safe the assumption that the entities expected to be found in them were players, coaches and teams and that the relations implied between them were the ones of current membership (i.e. players and coaches related to their current team). A similarly specialized scenario was in the project GLOCAL², involving the disambiguation of location entities within historical texts describing military conflicts. Again, the nature of these texts allowed us to expect to find in them, among others, military conflicts, locations where these conflicts took place and people and groups that participated in them.

Given that, in this paper we define a novel ontology-based disambiguation framework that is particularly applicable to similar to the above scenarios where knowledge about what entities and relations are expected to be present in the texts is available. Through a structured semi-automatic process the framework enables i) the exploitation of this a priori knowledge for the selection of the subset of domain semantic information that is optimal for the disambiguation scenario at hand, ii) the use of this subset for the generation of disambiguation evidence and iii) the use of this evidence for the disambiguation of entities within the scenario's texts. As we will show in the rest of the paper, this process allows our system to be more effective in such constrained scenarios than other disambiguation approaches designed to work in unconstrained ones.

The rest of the paper is as follows. Section 2 presents related work while section 3 describes in detail our proposed framework. Section 4 presents experimental results regarding the framework's effectiveness in the two application scenarios mentioned above. Finally, in sections 5 and 6 we make a critical discussion of our work, we summarize its key aspects and we outline the potential directions it could take in the future.

2 Related Work

A recent ontology-based entity disambiguation approach is described in [7] where an algorithm for entity reference resolution via Spreading Activation on RDF Graphs is proposed. The algorithm takes as input a set of terms associated with one or more ontology elements and uses the ontology graph and spreading activation in order to compute Steiner graphs, namely graphs that contain at least one ontology element for each entity. These graphs are then ranked according to some quality measures and the highest ranking graph is expected to contain the elements that correctly correspond to the entities.

Another approach is that of [4] where the application of restricted relationship graphs (RDF) and statistical NLP techniques to improve named entity annotation in challenging Informal English domains is explored. The applied restrictions are i) domain ones where various entities are a priori ruled out and ii) real world ones that can be identified using the metadata about entities as they appear in a particular post (e.g. that an artist has released only one album, or has a career spanning more than two decades).

² <http://glocal-project.eu/>

In [5] Hassel et al. propose an approach based on the DBLP-ontology which disambiguates authors occurring in mails published in the DBLP-mailing list. They use ontology relations of length one or two, in particular the co-authorship and the areas of interest. Also, in [12] the authors take into account the semantic data's structure, which is based on the relations between the resources and, where available, the human-readable description of a resource. Based on these characteristics, they adapt and apply two text annotation algorithms: a structure based one (Page Rank) and a content-based one.

Several approaches utilize Wikipedia as a highly structured knowledge source that combines annotated text information (articles) and semantic knowledge (through the DBPedia³ [1] and YAGO [13] ontologies). For example, DBPedia Spotlight [8] is a tool for automatically annotating mentions of DBPedia resources in text by using i) a lexicon that associates multiples resources to an ambiguous label and which is constructed from the graph of labels, redirects and disambiguations that DBPedia ontology has and ii) a set of textual references to DBPedia resources in the form of Wikilinks. These references are used to gather textual contexts for the candidate entities from wikipedia articles and use them as disambiguation evidence.

A similar approach that uses the YAGO ontology is the AIDA system [6] which combines three entity disambiguation measures: the prior probability of an entity being mentioned, the similarity between the contexts of a mention and a candidate entity, and the semantic coherence among candidate entities for all mentions together. The latter is calculated based on the distance between two entities in terms of type and subclassOf edges as well as the number of incoming links that their Wikipedia articles share.

The difference between the above approaches and our framework is detected in the way they treat the available semantic data. For example, Spotlight uses the DBPedia ontology only as an entity lexicon without really utilizing any of its relations, apart from the redirect and disambiguation ones. Thus, it's more text-based than ontology-based. On the other hand, AIDA builds an entity relation graph by considering only the type and subclassOf relations as well as "assumed" relations inferred by the links within the articles. The problem with this approach is that important semantic relations that are available in the ontology are not utilized and, of course, there is no control over which edges of the derived ontology graph should be utilized in the given scenario. Such control is not provided either in [7] or any of the rest aforementioned approaches except for that of [5] which, however, is specific for the scientific publications domain.

3 Proposed Disambiguation Framework

Our framework targets the task of entity disambiguation based on the intuition that a given ontological entity is more likely to represent the meaning of an ambiguous term when there are many ontologically related to it entities in the

³ <http://dbpedia.org>

text. These related entities can be seen as **evidence** whose quantitative and qualitative characteristics can be used to determine the most probable meaning of the term. For example, consider a historical text containing the term “Tripoli”. If this term is collocated with terms like “*Siege of Tripolitsa*” and “*Theodoros Kolokotronis*” (the commander of the Greeks in this siege) then it is fair to assume that this term refers to the city of Tripoli in Greece rather than the capital of Libya.

Nevertheless, as we already showed in the introduction, which entities and to what extent should serve as evidence in a given scenario depends on the domain and expected content of the texts that are to be analyzed. For that, the key ability our framework provides to its users is to construct, in a semi-automatic manner, semantic evidence models for specific disambiguation scenarios and use them to perform entity disambiguation within them. In particular, our framework comprises the following components:

- A **Disambiguation Evidence Model** that contains, for a given scenario, the entities that may serve as disambiguation evidence for the scenario’s target entities (i.e. entities we want to disambiguate). Each pair of a target entity and an evidential one is accompanied by a degree that quantifies the latter’s evidential power for the given target entity.
- A **Disambiguation Evidence Model Construction Process** that builds, in a semi-automatic manner, a disambiguation evidence model for a given scenario.
- An **Entity Disambiguation Process** that uses the evidence model to detect and extract from a given text terms that refer to the scenario’s target entities. Each term is linked to one or more possible entity uris along with a confidence score calculated for each of them. The entity with the highest confidence should be the one the term actually refers to.

In the following paragraphs we elaborate on each of the above components.

3.1 Disambiguation Evidence Model and its Construction

For the purposes of this paper we define an ontology as a tuple $O = \{C, R, I, i_C, i_R\}$ where

- C is a set of concepts.
- I is a set of instances.
- R is a set of binary relations that may link pairs of concept instances.
- i_C is a concept instantiation function $C \rightarrow I$.
- i_R is a relation instantiation function $R \rightarrow I \times I$.

The **Disambiguation Evidence Model** defines for each ontology instance which other instances and to what extent should be used as evidence towards its correct meaning interpretation. More formally, given a domain ontology O , a disambiguation evidence model is defined as a function $dem : I \times I \rightarrow [0, 1]$. If

$i_1, i_2 \in I$ then $dem(i_1, i_2)$ is the degree to which the existence, within the text, of i_2 should be considered an indication that i_1 is the correct meaning of any text term that has i_1 within its possible interpretations.

To construct the optimal evidence model for a given disambiguation scenario we proceed as follows: First, based on the scenario, we determine the concepts the instances of which we wish to disambiguate (e.g. players, teams and managers for the football match scenario). Then, for each of these concepts, we determine the related to them concepts whose instances may serve as contextual disambiguation evidence. The result of the above analysis should be a disambiguation evidence concept mapping function $ev_C : C \rightarrow C \times R^n$ which given a target concept $c_t \in C$ returns the concepts which may act as evidence for it along with the ontological relations whose composition links this concept to the target one. Table 1 contains an example of such a function for the football match descriptions scenario where, for instance, soccer players provide evidence for other soccer players that play in the same team. This mapping, shown in the second row of the table, is facilitated by the composition of the relations **dbpprop:currentclub** (that relates players to their current teams) and its inverse one **is dbpprop:currentclub of** (that relates teams to their current players). Table 2 illustrates a similar mapping for the military conflict texts scenario.

Table 1. Sample Disambiguation Evidence Concept Mapping for Football Match Descriptions

Target Concept	Evidence Concept	Relation(s) linking Evidence to Target
dbpedia-owl:SoccerPlayer	dbpedia-owl:SoccerClub	is dbpprop:currentclub of
dbpedia-owl:SoccerPlayer	dbpedia-owl:SoccerPlayer	dbpprop:currentclub, is dbpprop:currentclub of
dbpedia-owl:SoccerClub	dbpedia-owl:SoccerPlayer	dbpprop:currentclub
dbpedia-owl:SoccerClub	dbpedia-owl:SoccerManager	dbpedia-owl:managerClub
dbpedia-owl:SoccerManager	dbpedia-owl:SoccerClub	is dbpedia-owl:managerClub of

Using the disambiguation evidence concept mapping, we can then automatically derive the disambiguation evidence model dem as follows: Given a target concept $c_t \in C$ and an evidence concept $c_e \in C$ then for each instance $i_t \in i_C(c_t)$ and $i_e \in i_C(c_e)$ that are related to each other through the composition of relations $\{r_1, r_2, \dots, r_n\} \in ev_C(c_t)$ we derive the set of instances $I_t \subseteq I$ which share common names with i_t and are also related to i_e through $\{r_1, r_2, \dots, r_n\} \in ev_C(c_t)$. Then the value of dem for this pair of instances is computed as follows:

$$dem(i_t, i_e) = \frac{1}{|I_t|} \quad (1)$$

Table 2. Sample Disambiguation Evidence Concept Mapping for Military Conflict Texts

Target Concept	Evidence Concept	Relation(s) linking Evidence to Target
dbpedia-owl:PopulatedPlace	dbpedia-owl:MilitaryConflict	dbpprop:place
dbpedia-owl:PopulatedPlace	dbpedia-owl:MilitaryConflict	dbpprop:place, dbpedia-owl:isPartOf
dbpedia-owl:PopulatedPlace	dbpedia-owl:MilitaryPerson	is dbpprop:commander of, dbpprop:place
dbpedia-owl:PopulatedPlace	dbpedia-owl:PopulatedPlace	dbpedia-owl:isPartOf
dbpedia-owl:MilitaryPerson	dbpedia-owl:MilitaryConflict	dbpprop:commander

The intuition behind this formula is that the evidential power of a given entity is inversely proportional to the number of different target entities it provides evidence for. If, for example, a given military person has fought in many different locations with the same name, then its evidential power for this name is low.

3.2 Entity Disambiguation Process

The entity reference resolution process for a given text document and a disambiguation evidence model starts by extracting from the text the set of terms T that match to some instance belonging to a target or an evidence concept, that is some $i \in i_C(c)$, $c \in C_t \cup C_e$. Along with that we derive a term-meaning mapping function $m : T \rightarrow I$ that returns for a given term $t \in T$ the instances it may refer to. We also consider I_{text} to be the superset of these instances.

Then we consider the set of potential target instances found within the $I_{text}^t \subseteq I_{text}$ and for each $i_t \in I_{text}^t$ we derive all the instances i_e from I_{text} for which $dem(i_t, i_e) > 0$. Subsequently, by combining the evidence model dem with the term meaning function m we are able to derive an entity-term support function $sup : I_{text}^t \times T \rightarrow [0, 1]$ that returns for a target entity $i_t \in I_{text}^t$ and a term $t \in T$ the degree to which t supports i_t :

$$sup(i_t, t) = \frac{1}{|m(t)|} \sum_{i_e \in m(t)} dem(i_t, i_e) \quad (2)$$

Using this function we are able to calculate for a given term in the text the confidence that it refers to the entity $i_t \in m(t)$ as follows:

$$conf(i_t) = \frac{\sum_{t \in T} K(i_t, t)}{\sum_{i'_t \in m(t)} \sum_{t \in T} K(i'_t, t)} * \sum_{t \in T} sup(i_t, t) \quad (3)$$

where $K(i_t, t) = 1$ if $sup(i_t, t) > 0$ and 0 otherwise. In other words, the overall support score for a given candidate target entity is equal to the sum of the

entity’s partial supports (i.e. function *sup*) weighted by the relative number of terms that support it. It should be noted that in the above process we adopt the one referent per discourse approach which assumes one and only one meaning for a term in a discourse.

4 Framework Application and Evaluation

To evaluate the effectiveness of our framework we applied it in the two scenarios we mentioned in the introduction, the one involving disambiguation in football match descriptions and the other in texts describing military conflicts. In both cases we used DBPedia as a source of semantic information and we i) defined a disambiguation evidence model for each scenario and ii) used these models to perform entity disambiguation in a representative set of texts. Then we measured the precision and recall of the process. Precision was determined by the fraction of correctly interpreted terms (i.e. terms for which the interpretation with the highest confidence was the correct one) to the total number of interpreted terms (i.e. terms with at least one interpretation). Recall was determined by the fraction of correctly interpreted terms to the total number of annotated terms in the input texts. It should be noted that all target terms for disambiguation in the input texts were known to the knowledge base (i.e. DBPedia).

Finally, the results of the above evaluation process were compared to those achieved by two publicly available semantic annotation and disambiguation systems, namely DBPedia Spotlight ⁴ [8], AIDA⁵ [6]. The two systems were chosen for comparison because i) they also use DBPedia as a knowledge source and ii) they provide some basic mechanisms for constraining the types of entities to be disambiguated, though not in the same methodical way as our framework does. Practically, the two systems merely provide the users the capability to select the classes whose instances are to be included in the process. In all cases, it should be made clear that the goal of this comparison was not to disprove the effectiveness and value of these systems as tools for open domain and unconstrained situations but rather to verify our claim that our approach is more appropriate for disambiguation in “controlled” scenarios, i.e. scenarios in which a priori knowledge about what entities and relations are expected to be present in the text is available. A useful evaluation of popular semantic entity recognition systems for open scenarios may be found at [11].

4.1 Football Match Descriptions Scenario

In this scenario we had to semantically annotate a set of textual descriptions of football match highlights like the following: *“It’s the 70th minute of the game and after a magnificent pass by Pedro, Messi managed to beat Claudio Bravo. Barcelona now leads 1-0 against Real.”*. These descriptions were used as meta-data of videos showing these highlights and our goal was to determine, in an

⁴ <http://dbpedia-spotlight.github.com/demo/index.html>

⁵ <https://d5gate.ag5.mpi-sb.mpg.de/webaida/>

unambiguous way, which were the participants (players, coaches and teams) in each video. The annotated descriptions were then to be used as part of a semantic search application where users could retrieve videos that showed their favorite player or team, with much higher accuracy.

To achieve this goal, we applied our framework and we built a disambiguation evidence model, based on DBPedia, that had as an evidence mapping function that of table 1. This function was subsequently used to automatically calculate (through equation 1) the function *dem* for all pairs of target and evidence entities. Table 3 shows a small sample of these pairs where, for example, Getafe acts as evidence for the disambiguation of Pedro Leon because the latter is a current player of it. Its evidential power, however, for that player is 0.5, since in the same team there is another player with the same name (i.e. Pedro Rios Maestre).

Table 3. Examples of Target-Evidential Entity Pairs for the Football Scenario

Target Entity	Evidential Entity	dem
dbpedia:Real_Sociedad	dbpedia:Claudio_Bravo_(footballer)	1.0
dbpedia:Pedro_Rodriguez_Ledesma	dbpedia:FC_Barcelona	1.0
dbpedia:Pedro_Leon	dbpedia:Getafe_CF	0.5
dbpedia:Pedro_Rios_Maestre	dbpedia:Getafe_CF	0.5
dbpedia:Lionel_Messi	dbpedia:FC_Barcelona	1.0

Using this model, we applied our disambiguation process in 50 of the above texts, all containing ambiguous entity references. The overall number of references was 126 with about 90% of them being ambiguous. In average, each ambiguous entity reference had 3 possible interpretations with player names being the most ambiguous. Table 4 shows the results achieved by our approach as well as by DBPedia Spotlight and AIDA. It should be noted that when using the latter systems, we used their concept selection facilities in order to constrain the space of possible interpretations. Still, as one can see from the table data, the constraining of the semantic data that our custom disambiguation evidence model facilitated (e.g. the consideration of only the current membership relation between players and teams) was more effective and managed to yield significantly better results.

Table 4. Entity Disambiguation Evaluation Results in the Football Scenario

System/Approach	Precision	Recall	F_1 Measure
Proposed Approach	84%	81%	82%
AIDA	62%	56%	59%
DBPedia Spotlight	85%	26%	40%

4.2 Military Conflict Texts Scenario

In this scenario our task was to disambiguate location references within a set of textual descriptions of military conflicts like the following: *“The Siege of Augusta was a significant battle of the American Revolution. Fought for control of Fort Cornwallis, a British fort near Augusta, the battle was a major victory for the Patriot forces of Lighthorse Harry Lee and a stunning reverse to the British and Loyalist forces in the South”*. For that we used again DBPedia and we defined the disambiguation evidence mapping function of table 2 which, in turn, produced the evidence model that is (partially) depicted in table 5.

Table 5. Examples of Target-Evidential Entity Pairs for the Military Conflict Scenario

Location	Evidential Entity	dem
dbpedia:Columbus,_Georgia	James H. Wilson	1.0
dbpedia:Columbus,_New Mexico	dbpedia:Pancho_Villa	1.0
dbpedia:Beaufort_County,_South_Carolina	dbpedia:Raid_at_Combahee_Ferry	1.0
dbpedia:Beaufort_County,_South_Carolina	dbpedia:James_Montgomery_(colonel)	1.0
dbpedia:Beaufort_County,_North_Carolina	dbpedia:Battle_Of_Washington	1.0
dbpedia:Beaufort_County,_North_Carolina	dbpedia:John_G._Foster	1.0

Using this model, we applied, as in the football scenario, our disambiguation process in a set of 50 military conflict texts, targeting the locations mentioned in them. The average reference ambiguity of this set was 5 in a total of 55 locations. Table 6 shows the achieved results which verify the ability of our framework to improve disambiguation effectiveness.

Table 6. Entity Disambiguation Evaluation Results in the Military Conflict Scenario

System/Approach	Precision	Recall	F_1 Measure
Proposed Approach	88%	83%	85%
DBPedia Spotlight	71%	69%	70%
AIDA	44%	40%	42%

5 Discussion

It should have been made clear from the previous sections that our framework is not independent of the content or domain of the input texts but rather adaptable to them. That’s exactly its main differentiating feature as our purpose was not to build another generic disambiguation system but rather a reusable framework that can i) be relatively easily adapted to the particular characteristics of the domain and application scenario at hand and ii) exploit these characteristics in

order to increase the effectiveness of the disambiguation process. Our motivation for that was that, as the comparative evaluation of the previous section showed, the scenario adaptation capabilities of existing generic disambiguation systems can be inadequate in certain scenarios (like the ones described in this paper), thus limiting their applicability and effectiveness.

Of course, the usability and effectiveness of our approach is directly proportional to the content specificity of the texts to be disambiguated and the availability of a priori knowledge about their content. The greater these two parameters are, the more applicable is our approach and the more effective the disambiguation is expected to be. The opposite is true as the texts become more generic and the information we have out about them more scarce. A method that could a priori assess how suitable is our framework for a given scenario would be useful, but it falls outside the scope of this paper. Also, the framework's approach is not completely automatic as it requires some knowledge engineer or domain expert to manually define the scenario's disambiguation evidence mapping function. Nevertheless, this function is defined at the schema level thus making the number of required mappings for most scenarios rather small and manageable.

Finally, although we haven't formally evaluated the scalability of our approach, the fact that our framework is based on the constraining of the semantic data to be used makes us expect that it will perform faster than traditional approaches that use the whole amount of data. Furthermore, as the disambiguation evidence model may be constructed offline and stored in some index, the most probable bottleneck of the process will be the phase of determining the candidate entities for the extracted terms rather than the resolution process. Nevertheless, a more rigorous scalability study will have to be made as part of future work.

6 Conclusions and Future Work

In this paper we proposed a novel framework for optimizing named entity disambiguation in well-defined and adequately constrained scenarios through the customized selection and exploitation of semantic data. First we described how, given a priori knowledge about the domain(s) and expected content of the texts that are to be analyzed, one can use the semantic data and define an evidence model that determines which and to what extent semantic entities should be used as contextual evidence for the disambiguation task at hand. Then we described the process through which such a model can be actually used for this task. The overall framework was experimentally evaluated in two specific scenarios and the results verified its superiority over existing approaches that are designed to work in open domains and unconstrained scenarios.

Future work will focus on the further automation of the disambiguation evidence model construction by means of data mining and machine learning techniques. Moreover, an online tool to enable users to dynamically build such models out of existing semantic data and use them for disambiguation purposes, will be developed.

Acknowledgements

This work was supported by the Spanish project CENIT-2009-1026 BuscaMedia and by the European Commission under contract FP7- 248984 GLOCAL.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann J., Cyganiak, R., Ives, Z.G.: DBpedia: A Nucleus for a Web of Open Data. In Proceedings of the 6th International Semantic Web Conference, pages 722-735, 2007.
2. Fader, A., Soderland, S., Etzioni, O.: Scaling wikipedia-based named entity disambiguation to arbitrary web text. In Proceedings of the WikiAI 09 - IJCAI Workshop: User Contributed Knowledge and Artificial Intelligence: An Evolving Synergy, Pasadena, CA, USA, July 2009.
3. Ferragina, P., Scaiella, U.: TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, New York, SA, 1625-1628.
4. Gruhl, D., Nagarajan, M., Pieper, J., Robson, C., Sheth A.P.: Context and domain knowledge enhanced entity spotting in informal text. In Proceedings of the 8th International Semantic Web Conference, pages 260-276, 2009.
5. Hassell, J., Aleman-Meza, B., Arpinar, I.: Ontology-driven automatic entity disambiguation in unstructured text. In Proceedings of the 3rd European Semantic Web Conference, pages 44-57, Springer Berlin, Heidelberg, 2006.
6. Hoffart, J., Yosef, M.A., Bordino, I., Frstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 782-792.
7. Kleb, J., Abecker, A.: Entity Reference Resolution via Spreading Activation on RDF-Graphs. In Proceedings of the 7th European Semantic Web Conference, pages 152-166, Springer Berlin, Heidelberg, 2006.
8. Mendes, P.N., Jakob, M., Garcia-Silva, A., Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems, ACM, New York, USA, 1-8, 2011.
9. Miller, G., Charles, W.: Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):128.
10. Pilz, A., Paass, G.: Named entity resolution using automatically extracted semantic information. *Workshop on Knowledge Discovery, Data Mining, and Machine Learning*, page 84-91, 2009.
11. Rizzo G., Troncy, R.: NERD: A Framework for Evaluating Named Entity Recognition Tools in the Web of Data. In 10th International Semantic Web Conference, Demo Session, pages 1-4, Bonn, Germany, 2011.
12. Rusu, D., Fortuna, B., Mladenic, D.: Automatically Annotating Text with Linked Open Data. In 4th Linked Data on the Web Workshop (LDOW 2011), 20th World Wide Web Conference, Hyderabad, India, 2011.
13. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Core of Semantic Knowledge. In 16th World Wide Web Conference, 2007.