Market Intelligence: Linked Data-driven Entity Resolution for Customer and Competitor Analysis

Ulli Waltinger[1], Dan Tecuci[2], Florin Picioroaga[3], Cosmin Grigoras[3], and Sean Sullivan[4]

[1] Siemens AG Corporate Technology, Munich, Germany
[2] Siemens Corporation, Corporate Technology Princeton, NJ, USA
[3] Siemens AG Corporate Technology, Brasov, Romania
[4] Siemens Energy Inc. Orlando, USA
{ulli.waltinger,dan.tecuci,florin.picioroaga,
cosmin.grigoras,sean.sullivan}@siemens.com
http://www.siemens.com/

Abstract. In this paper, we present a linked data-driven method for named entity recognition and disambiguation which is applied within an industry customer and competitor analysis application. The proposed algorithm primarily targets the domain of geoparsing and geocoding, but it can easily be adapted to other problems such duplicate detection. The contributions of this paper are three fold: First, we want to give an overview of *Market Intelligence*, a customer and competitor analysis application developed for Siemens Energy, which allows users to pose questions and queries on regularly crawled websites, emails and RSS feeds, to detect and respond to competitor, customer, and market trends more effectively. Second, we describe the UIMA-based processing architecture that builds the framework for analyzing and converting unstructured heterogeneous documents into a structured and semantically-enhanced knowledge representation. Third, we propose a novel algorithm that is used within the framework for content analysis and entity disambiguation. The performed evaluation shows with an accuracy of up to 91.69% that the proposed method for named entity recognition and disambiguation is very effective, while at the same time relying on Linked Data only.

Keywords: Named Entity Recognition, Named Entity Disambiguation, Word Sense Disambiguation, GeoParsing, GeoCoding, Market Intelligence.

1 Introduction

Today enterprises deal with decisions that involve the analysis of information from various heterogeneous sources on a massively scale. In this context, an effective information access and analysis can be seen as one of the fundamental building blocks within the decision making process and in the process of

F. Daniel, P. Dolog, and Q. Li (Eds.): ICWE 2013, LNCS 7977, pp. 467–481, 2013.

enabling an cost-effective customer service [1]. The amount of available information nowadays grows at an amazing speed, which raises several challenges. More precisely, it is assumed that enterprise data will grow by 800 percent in the next five years, whereas 80 percent of it exists by means of documents, files or other unstructured data [1]. That is, most of the data and resources available lack of meta data or being semantically augmented, which support an efficient and well-defined data exploration and analysis for market intelligence applications. In this context information extraction and retrieval techniques, such as Named Entity Recognition (NER) and Disambiguation (NERD) [2, 3], are an important part for obtaining and automatically analyzing such information hidden in unstructured, machine-readable documents. Especially in the area of customer and competitor analysis applications, plays the automatic identification and resolution of entities, such as company names, their location and connected profiles a significant role. These applications aim to provide information about business opportunities, strengths and weaknesses of customers/competitors that are primarily distributed across various unstructured sources. In the setup of *Market Intelligence*, a project of Siemens Corporation, Corporate Technologies and Siemens Energy, we aim to identify customer and competitor information from unstructured documents to enable answers such as: `What are the service units of company X that are located around Clive and Jupiter?`, `Which units on the East Coast remain open?` or `Is there a company X that has installed component Y?` In this context, the automatic extraction and disambiguation of context-specific entities and its geo-related references [4] are in the center of the project scope. That is, in this paper, we do not focus on the aspect of natural language question answering, but targeting the challenge of not only extracting business relevant information out of regularly crawled websites, emails and RSS feeds, but also applying a context-specific disambiguation, of the extracted information. As an example:

> "The units located in **Jupiter** [↦Jupiter, Florida] and **Princeton** [↦Princeton, British Columbia] (Canada) [↦Canada] will remain open."

> "**Princeton** [↦Princeton, New Jersey] (US) [↦United States], the city **New York** [↦New York City] and the state **New York** [↦New York] need to be notified."

In this example, one can identify that the surface form of the geo-related entities entities are often ambiguous. That is, taken out of context, the same name (e.g. *Princeton*, *Jupiter*, or *New York*) may have multiple meanings (i.e. refer to different entities). There are three main contributions of this paper: In Section 2, we give an overview of *Market Intelligence*, the customer and competitor analysis application developed for Siemens Energy, in which the described components are integrated. Section 3 reviews related work. Thereupon in Section 4, we describe the UIMA-based processing architecture that builds the framework for analyzing and converting unstructured heterogeneous documents into a structured and semantically-enhanced knowledge representation. In Section

5, we propose a NERD algorithm that is used within the framework for content analysis and entity disambiguation targeting the domain of geoparsing and geocoding. In Section 6, we present the evaluation of the entity disambiguation algorithm that is applied on two different datasets. Finally, Section 7 concludes this paper.

2 Overview of Market Intelligence

Information about business opportunities, new regulations, competitor and customer news is massive and scattered across an ever growing number of sources. Hidden in publicly available news, internal bulletins, market reports or documents it is difficult to keep track of latest developments and get a global picture of the market situation. The goal of the *Market Intelligence* application is to aggregate and analyze such data and extract actionable knowledge from it. Currently we focus on the following functionalities: automated classification of incoming information into business relevant categories, automatic identification of named entities from a catalog of entities of interest, instant notification based on custom-made rules that use the result of classification and entity recognition, and collaboration (sharing and commenting). The data that is analyzed comes from a set of publicly available websites identified by the business as being of interest. Among the named entities identified, geolocations are of great importance. Figure 1 shows a map representation of a set of selected news and Figure 2 shows an individual piece of news with its corresponding annotations.

The application is being developed for Siemens Energy Service.

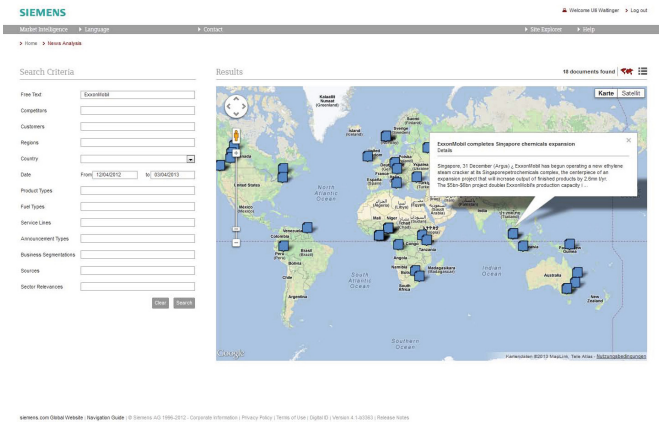

Fig. 1. Screenshot of the Market Intelligence application

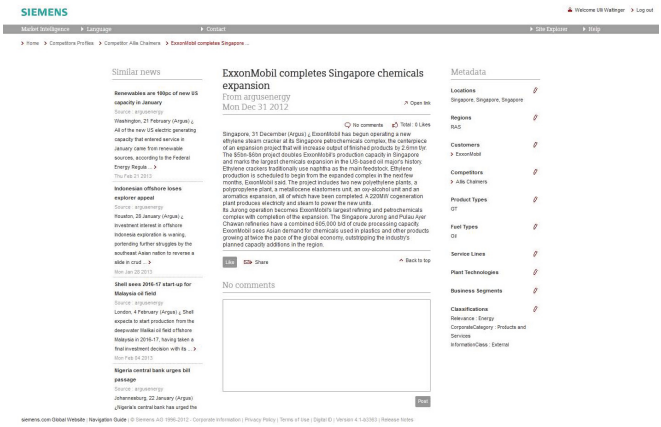

Fig. 2. Metadata annotations in Market Intelligence web application

3 Related Work

Documents, articles and other comprised resources contain named entities of different flavor, as for example locations, products or people, but also weapons or organizations, which play a significant role in automatic data analytic (e.g. product and relationship mining, location detection, sentiment analysis). Named Entity Recognition (NER) has been extensively addressed in different research fields [5, 6, 3], and can be seen as one of the fundamental components of current information extraction and retrieval sytems. This task focuses on the identification of proper nouns, which are further classified into a predefined set of entity categories (e.g. location, persons, numeric or time). As an extension of it, the task of Named Entity Disambiguation (NED), attempts to additionally disambiguate the classified entity by linking the entity to real world object identifier (e.g. URI). That is, mapping information units to explicitly and uniquely mentioned entities in a knowledge base. As one of the most prominent comprised knowledge base, the *Wikipedia* data set was heavily used for this task lately [7, 6]. In this context, numerous approaches have been published using concept similarity [8–11] or relatedness measures [6, 12] to rank the respective object candidates.

Most recently, RDF-based knowledge bases such as *Freebase*[1], *GeoNames*[2], *YAGO* [13], or *DBpedia* [14] are used as a resource for web-based entity identifier [15, 16]. For an comprehensive overview and comparison of current (publicly available) NERD services that leverage RDF-based repositories as a resource see [3]. The domain of geocoding or geoparsing [4], can be seen as a geospatial extension of NER(D). This research field is concerned with the automatically mapping of locations specifically, referred to as the processing of textually-encoded spatial

[1] `www.freebase.com`
[2] `www.geonames.org`

data [17]. Note that we see geoparsing as the task of location-based extraction from text (NER), and gecoding as the NED complement, the mapping of references to real-world counterparts [17]. Similar to current NER approaches, we can identify three different branches of methods [4]: Gazetteer-based lookup methods [18], Rule-based approaches (e.g. GATEs ANNIE module [19]) by using a set of symbolic rules to encode the decision procedure (Definite Clause Grammars via Prolog) [20]. The third branch uses machine learning-based approaches. Most commonly using a sliding window, which is introduced in order to extract a set of classification properties and features (e.g. context, length, string surface) [21]. In this work, we are using for the NER component, the Gazetteer-based approach as a stimulus for the learning-based entity classification. With regards to NED, we are focusing on meta data via *DBpedia* only. In this context, the approaches of [10] and [16] are most related to our approach in the sense of putting the textual context of an entity in the center for the task of candidate ranking. However, different to others, our approach does not rely on any training cycle for graph construction or edge weighting, but operates entirely on the RDF-metadata only. In addition, the method proposed in this paper allows to incorporate the (initial) named entity category as an stimulus and part of the evidence strategy for disambiguation.

4 Information Processing Architecture

The overall processing architecture of the *Market Intelligence* application can be divided into two interconnect pipelines: the data management pipeline, which leads the data workflow between the different processing components, and the UIMA pipeline that bears the content extraction and analysis procedures.

4.1 Data Management Pipeline

The process of data transformation consists of the following components (see Figure 3):

1. **Content Dispatcher:** This components collects the heterogeneous data from various sources (RSS, E-Mails, crawled web pages) via custom adapters.
2. **Content Transformer:** The aggregated data collection gets pre-processed via processing templates (i.e. extracting the only the text from the incoming data).
3. **Content Storage:** The extracted content fragments are stored to a relational database storage.
4. **Message Broker:** Each content fragment is further sent out as a message to a message broker.
5. **UIMA Connector:** The message from the message broker is consumed by a component (pipeline connector) responsible for transforming the received data to a data structure (Common Analysis Structure) accepted by the UIMA framework.

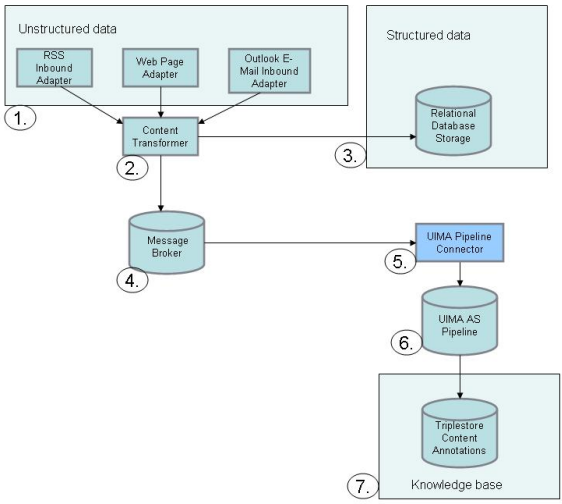

Fig. 3. Overview of the general processing work flow of the Marketing Intelligence pipeline

6. **UIMA Pipeline:** UIMA pipeline runs the analysis engines on the given data and extracts the information specific to each engine. We will call this information annotation.
7. **Annotation Storage:** The extracted annotations are stored in a knowledge base as triple statements.

4.2 UIMA Pipeline

The *UIMA framework*[3] has been used for analyzing the text message and extracting the information required. More precisely, the framework consists of a set of text analyzing engines that are grouped in a single processing pipeline. The analyzing engines are refereed to as *annotators*. The results of these *annotators* are defined as *annotations*. All these *annotations* are then persisted by a special component called *consumer*. The *Market Intelligence* project incorporates the following UIMA components:

1. **LocationAnnotator** that recognizes the geographical locations (cities, countries) that occur in the respective content fragments. This component additionally resolves each entity by its unique URI representation (see Section 5).
2. **OrganizationAnnotator** that recognizes customers and competitors entities. It utilizes the gazetteer component as available within *GATE*[4].

[3] `http://uima.apache.org/`
[4] `http://gate.ac.uk/`

3. **ClassifierAnnotator** that recognizes domain-specific meta-information, such as *fuel type*, *business segmentation*, *joint venture* etc. using Support Vector Machines.
4. **RegularExpressionAnnotator** that is used for matching meta-information, based on a set of regular expressions.
5. **RDFCASConsumer** is used to store the annotations in the RDF-based triple format.

Subsequently, all annotations produced by components within the UIMA pipeline can be viewed and further processed (deleted or adding new ones) within the Market Intelligence web application.

5 Evidence-Based Entity Disambiguation

As described in the previous section, the UIMA pipeline integrates several entity annotators and an entity disambiguation annotator that focuses on geo-related references. The work flow of this evidence-based component is depicted in Figure 4, and can be subdivided into three consecutive components. At first, the recognition task that combines a state-of-the-art NER library with domain-specific gazetteer induction. Second, the disambiguation task which utilizes *DBpedia* as an resource for entity disambiguation and URI identifier assignment. Finally, the connector to the MI application, which makes use of the data set of *GeoNames* to construct SPARQL queries based on prior templates. In the following, we describe each individual component in more detail.

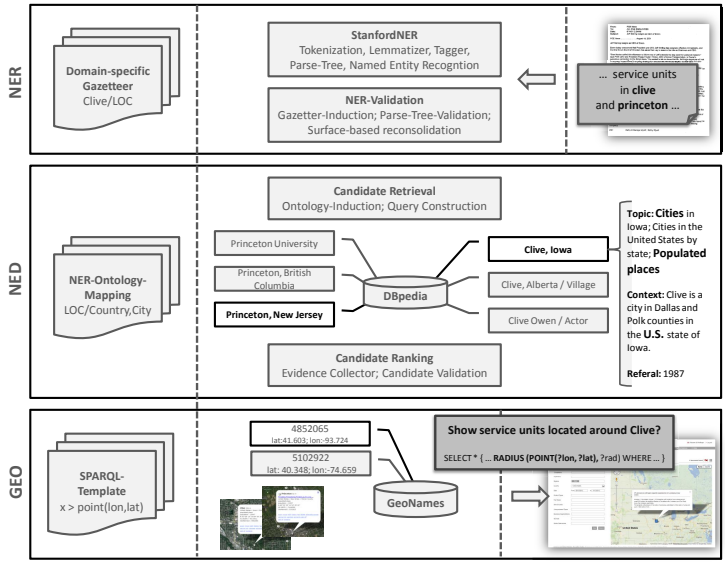

Fig. 4. Overview of the work flow of the evidence-based geospatial named entity recognition and disambiguation

5.1 NER - Named Entity Recognition

The NER phase focuses on the identification and tagging of single nouns and phrases using *StanfordNER* ([21]). That is, each input document is preprocessed by applying tokenization, lemmatization, part-of-speech-tagging, parse-tree extraction and (default) named entity recognition. Subsequently, a NER validation is applied that enhances and corrects the default token representation as generated by *StanfordNER*. This component utilizes domain-specific gazetteers (e.g. clive may also refer to a city) to re-annotate those text segments, which could not be identified within the pre-processing phase. In addition, multi-word units (e.g. jupiter, florida), which are missed by *StanfordNER* will be concatenated by reconciliation the surface- and the parse-tree representation. The resultant *StanfordNER*-enhanced object representation is further used as an input for the NED component.

5.2 NED - Named Entity Disambiguation

The NED components can be subdivided into the candidate retrieval and the candidate ranking module.

Candidate Retrieval. The candidate retrieval module utilizes the *DBpedia* data set not only to retrieve a a list of object identifier, but also uses the ontology[5] to typify and validate the search strategy. That is, we incorporate a mapping between the entity category (e.g. LOC for location) and the respective *DBpedia* counterparts as they are represented using the *SKOS* vocabulary (e.g. Country, City, ...).

$$LOC \mapsto PopulatedPlace; Geography; CelestialBody; NaturalPlace; \cdots \quad (1)$$

This changes massively the search strategy, since we are inducing higher confidence to candidates which are associated to a certain category (e.g. clive rather location than name), instead of the most common "eat-all-you-can" approach. In order to allow an efficient search-and-retrieval performance, we decided to parse the entire DBpedia data set into an *Apache Lucene*[6] index. Note, we used only a snapshot of the meta data (as represented through title, short abstracts, articles categories and the links to GeoNames) for index construction. For each entity candidate we construct the Ontology-induced query and score each entity as follows:

$$score_{can}(q,d) = \log_{10}(\sum_{t \in q}(tf(t \in d) \cdot idf(t)^2 \cdot t_b \cdot norm(t,d))) \quad (2)$$

where $tf(t \in d)$ defines the term frequency within the observed scored DBpedia short abstract description d; $idf(t)$ represents the inverse document frequency

[5] `http://wiki.dbpedia.org/Ontology`
[6] `http://lucene.apache.org`

applied to the DBpedia summary description representation. t_b is the search time boost of term t in the query q. $norm(t, s)$ encapsulates a few (indexing time) boosts and length factors with reference to *Lucene's* document and field boost property [22]. Note, we collect only the 100 best entity candidates which are further passed to the ranking component.

Candidate Ranking: The candidate ranking can be subdivided into the evidence collection and the evidence validation phase. More precisely, at first, we collect a number of evidences that consider the confidence and the probability that a candidate refers to a proper entity instance, in order to, subsequently, rank and validate the most likely referred instances of a given candidate. The evidence collector utilizes the following measures:

Popularity-Based Evidence: This measure follows the rational that given an surface form of an token there exist a prior assumption of which entity might be meant. As for instance, just given the context "*We live in New York*", the majority of people would think of the city rather than the state of *New York*. This prior stimulus can be deduced from the number of referels (or backlinks), which are interlinked to a certain entity. That is, the number of pointing hyperlinks, established by human, operates as a *common sense* amplifier for a certain entity, as proposed by [23]. Though, we define the popularity-based evidence score as follows:

$$evi_{pop}(u) = n \cdot \log_{10}(\log_{10}(b_u)) \tag{3}$$

where b_u refers to the number of incoming links to a certain DBpedia entity u. That is, we use the double logarithmic normalized backlink score as an (probability-based) evidence for the most popular referenced DBpedia instance for a given surface form (e.g. *Princeton (New Jersey)*: 0.51 v.s. *Princeton (British Columbia)*: 0.30)

Surface-Based Evidence: The surface-based evidence refers to quotient between the intersection and the union of the pairwise compared term features among the input entity, e, and the current observed DBpedia entity u (e.g. Princeton $\mapsto$ Princeton (New Jersey))

$$evi_{sur}(e, u) = \frac{tf_{e,u}}{tf_{e,u} + tf_e + tf_u} \tag{4}$$

That is, this score collects evidences for the term overlap on its surface form. While the surface-based evidence is a good indicator for a successful mapping, there exist a lot of false positive examples for it. For example: surface form of *Aspen* $\mapsto$ *Aspen* within *DBpedia* (evi_{sur} : 1.0), though the article *Aspen* describes not *Aspen, Colorado* but a certain tree species.

Context-Based Evidence: The context-based score collects evidence from the description of each *DBpedia* entity. The rational behind this evidence score is that each (surface form) of an entity is primarily instantiated through its context.

We define context as the surrounding terms co-occur (left/right) with the entity within a word window of size m.

$$evi_{con}(c,u) = \frac{\sum_{i=1}^{n} c_i \times u_i}{\sqrt{\sum_{i=1}^{n} (c_i)^2} \times \sqrt{\sum_{i=1}^{n} (u_i)^2}} \tag{5}$$

That is, we apply the standard cosine measure to obtain the similarity between the context of the input entity (c) using m word features left and right from the observed token by means of its sentence representation, and the context of the entity candidate as given by its short summary value (u). Note that we applied the normalized term frequency for input vector construction utilizing nouns and ner entities only.

Topic-Based Evidence: The topic-based evidence measures the correlation from the initial mapped named entity category and the respective *DBpedia* category associated to the candidate. More precisely, since we are able to traverse the category taxonomy within *DBpedia* by means of its graph-based representation (e.g. *Princeton (New Jersey)* $\mapsto$ *University towns in the United States* $\mapsto$ *Cities in the United States*), we are able to score the normalized graph-path distance, between the initial mapping category uc and the respective category candidate note nc by:

$$evi_{top}(uc,nc) = 1/|dis(uc,nc)| \tag{6}$$

The rational behind this approach is to allow to adjust the confidence of an entity candidate by its assigned category even if the latter was not part of the initial ontology mapping process (e.g. LOC $\mapsto$ Country, City,...). Note, we allowed also a substring match of nodes to score the distance (e.g. *City* $\mapsto$ *Cities in the United States*).

Mutual Evidence Confidence: In this NED phase, the k evidences are accumulated to a mutual confidence score defined as

$$conf_{evi}(d) = \frac{score_{can} + \sum_{i=1}^{k} (\lambda \cdot evi_i) + \phi}{k+2} \tag{7}$$

$$d_{max} = \arg\max_{d \in D} conf_{evi}(d) \tag{8}$$

where ϕ represents the redirect amplifier as an indicator whether an respective redirect instance was used for the calculation (e.g. *NYC* $\mapsto$ *New York City*). λ is defined as a weighting parameter (evaluation setup $\lambda = 2$). In a final step, all entity candidates, $d \in D$ are ranked by its $conf_{evi}$ score, and the final entity instance is selected by means of $d_{max} > \mu$. That is, we allow the disambiguation assignment only for those entities with a sufficient mutual evidence confidence ($\mu = 0.3$).

5.3 GEO - Geospatial Analysis

The final component in the processing pipeline is the geospatial analysis. Here, the newly assign *DBpedia* URI is mapped to it *GeoNames* URI counterpart. For this task, we use the already available triples linking within the *DBpedia* data set. Having successfully assigned a given *GeoNames* URI, we apply different *SPARQL* template queries to collect the information need for the MI applications. As for example to infer from a *city* $\mapsto$ *country*, or it's *geographic coordinates* as: $Clive_{raw} \mapsto Clive, UnitedStates_{nerd} \mapsto 4852065_{geo} \mapsto (lat : 41.60304; lon : -93.72411_{geo}) \mapsto State : Iowa_{geo} \mapsto Country : UnitedStates_{geo}$

The set of inferred geographical information is finally stored within the RDF-based triple store component, and subsequently get interlinked to the *GeoNames* dataset, and to business-related entities. The business entities are imported from the database by use of a translator importer, which requires the hidden semantics of the tabular form information to be declared up front and will be used for clusters of same type information.

6 Experiments

We conducted two different experiments, in order to evaluate the proposed evidence-based method to named entity recognition and disambiguation for the domain of geoparsing and geocoding. We decided to use two different data sets, not only to allow a generic comparison to other state-of-the-art approaches, but also to evaluate *both sides* of the targeted application. More precisely, since the algorithm is part of the pre-processing component of the MI application, it is used to facilitate both, not only the annotation process for the unstructured document collection, but also for the analysis of the questions and queries as posed by the users against MI application. Therefore, we decided to use for the evaluation of the backend side, a standard data set - the *CoNLL* task data set [24] - since multiple approaches have already been evaluated. The second data set, targets the user perspective of the application and is part of an entity-biased question-answering corpus collection [25]. Both data sets have already been manually annotated and build therefore the reference plain and bench mark for our evaluation. In the following, we describe the respective corpus properties in more detail.

6.1 Dataset

The *CoNLL* dataset was created by [16] based on the *CoNLL 2003* data [24]. It consists of 1.393 news article, which were manually annotated by means of corresponding *YAGO*[7] entities. Each of the total $34,956$ mentions was disambiguated by two students, with an overall distribution of 25 entities per article on average. For the experiment, we have used both, the test set with $4,458$ entities, refereed to as *CoNLL-TestA* and the training set with $27,790$ entities, denoted

[7] `http://www.mpi-inf.mpg.de/yago-naga/yago/`

as *CoNLL-TestB*. Note, since our approach does not rely on the existence of a training corpus, we used both within our experiments. For a detail description on the used corpus see [16]. The *QA* data set[8] consists of 5,500 questions initially created by [26] in the context of question classification. This question collection has been additionally processed by [25] comprising a training set of 5000 and a test set of 500 questions. For each question, the named entities were annotated and classified by means of the standard categories person, location and organization. For the experiments, we extracted only those question, which refer to location-based entities and manually assigned and validated their corresponding *DBpedia* URI's.

6.2 Results

The results of the backend evaluation (CoNLL) are shown in Table 1. We have used the results published by [16] and [11] as our reference base line. As we can identify the performance of the location based disambiguation performs, with an accuracy of over 91%, equally well on both datasets. The overall macro precision is slightly under the best performing system, the micro precision however is outperforming the benchmark results. Though, with regards to the disambiguation of organizations, we could identify the limitations of our algorithm. Given just the entity candidate *Barcelona*, our method classifies it as an location, and subsequently disambiguates the candidate to the city *Barcelona* rather than to the soccer team *F. C. Barcelona*, which potentially could be identified by the broader context of the entire document. Note that we used a sentence-based context window for the experiments.

Table 1. Result of test set B [16] using 1392 documents and 4458 entities. (Organization which are tagged as locations e.g Barcelona but it is F.C. Barcelona;) with Competitor Results.

	NERD TestA (4458).	NERD TestB (27790).	ML-Ref [16]	Kulk [11]
Acc - LOC	**91.69**	**91.50**		
Acc - PER	78.6	86.7		
Acc - ORG	43.8/76.5	60.7/78.6		
MicroPrec	73.1	**84.1**	81.82	72.87
MacroPrec	71.3	79.6	81.91	76.74

The results of the second experiments are shown at Table 2. As a reference baseline, we have used the results published by [25] and the results of *Nlp-Geo*[9] [27]. Even though, the second data set is with a size of 200 annotated questions rather small, our systems performs, with an accuracy of 83%, very well on the task of location-based entity recognition and disambiguation.

[8] Accessible at `https://qa.l2f.inesc-id.pt/wiki/index.php/Resources`

[9] `http://code.google.com/p/nlp-geo/`

Table 2. Result of NERQ data set as provided by [25] using 200 questions ((Geo-Precision)85.04 (Geo-Recall)62.43)

	NERD QA (200)	Geo-NLP[27]	Supervised [25]
Acc - LOC	**83,91**	75,11	59,43

6.3 Discussion

Overall, the result of both experiments show that our evidence-based method performs on a very satisfying basis on both sides of the application pipeline. Analyzing the individual errors of the evaluation more closely, we can identify some systematic issues using the evidence-based method: First, using a sentence-based context window allows a sufficient level for most of the test cases, though, it does not consider the topic of entire documents. As for example, the occurrence of *cyprus* is correctly identified as an entity, but mapped to *Cyprus* as the country instead of *Cyprus_national_football_team*, which the article referred to. Similar examples, *New_Zealand* as the country instead of *New_Zealand_national_rugby_union_team*, or *Birmingham* as the country instead of *Birmingham_City_F.C.* Second, synonymous entities (in terms of redirects) have not been separately evaluated or resolved. That is, even if the redirects within *DBpedia* map *World Wide Web* to *Internet* or *Islam* $\mapsto$ *Muslim*, we treated the assignment of *World Wide Web* as an error, if in the test set the entity *Internet* was used. Third, the algorithms makes use of the actual entity category as a stimulus for the disambiguation task. More precisely, different to other approaches, our method doe not disregard the entity category (e.g. LOC) during the disambiguation phase. However, wrongly classified entity information is passed on to the NED component, influencing the candidate retrieval and the topic-based evidence score. From the perspective of the access to the comprised Linked Data resources, for performance reasons, we decided to index the RDF-dataset of *DBpedia* in an offline mode. That is, we used a snapshot of selected metdata information to allow an efficient search-and-retrieval process. Though, this task could be also achieved using an endpoint service only, as it is deployed for the task of the *GeoNames* mapping.

7 Conclusion

In this paper, we gave an overview of the customer and competitor analysis application *Market Intelligence* for *Siemens Energy*. This system allows users to pose questions and queries on regularly crawled document repositories, to detect and respond to competitor, customer, and market trends more effectively. We described the overall UIMA-based processing architecture that builds the framework for analyzing and converting unstructured heterogeneous documents into a structured and semantically-enhanced knowledge representation. Finally, we presented a multiple evidence-based method for named entity recognition and disambiguation which is applied within the industry-based analysis application.

The proposed algorithm primarily targeted the domain of geoparsing, though, it's application was also evaluated for the domain of person and organization resolution. The performed evaluation shows with an accuracy of up to 91.69% that the proposed method for named entity recognition and disambiguation is very effective, while at the same time relying on *Linked Data* only.

References

1. IBM-Whitepaper, I.: Leveraging content integration for improved customer service. Technical report (2010)
2. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 100–110 (1999)
3. Rizzo, G., Troncy, R., Hellmann, S., Bruemmer, M.: NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In: 5th Workshop on Linked Data on the Web, LDOW, Lyon, France (April 16, 2012)
4. Hill, L.L.: Georeferencing: The Geographic Associations of Information. Digital Libraries and Electronic Publishing (2006)
5. Extracting company names from text. In: Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications, vol. i (1991)
6. Milne, D.N., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, pp. 509–518 (2008)
7. Bunescu, R., Pasca, M.: Using Encyclopedic Knowledge for Named Entity Disambiguation. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006) (2006)
8. Cucerzan, S.: Large-scale named entity disambiguation based on wikipedia data. In: Proceedings of the EMNLP-CoNLL, Prague, Czech Republic, June 28-30, pp. 708–716 (2007)
9. Nguyen, H.T., Cao, T.H.: Named entity disambiguation on an ontology enriched by Wikipedia. In: RIVF, pp. 247–254. IEEE (2008)
10. Waltinger, U., Mehler, A.: Who is it? context sensitive named entity and instance recognition by means of wikipedia. In: 2008 IEEE / WIC / ACM International Conference on Web Intelligence, WI 2008, Sydney, NSW, Australia, December 9-12. Main Conference Proceedings, pp. 381–384 (2008)
11. Kulkarni, S., Singh, A., Ramakrishnan, G., Chakrabarti, S.: Collective annotation of wikipedia entities in web text. In: Proceedings of the 15th ACM SIGKDD, KDD 2009, pp. 457–466. ACM, New York (2009)
12. Waltinger, U., Mehler, A.: Social semantics and its evaluation by means of semantic relatedness and open topic models. In: 2009 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2009, Milan, Italy, September 15-18. Main Conference Proceedings, pp. 42–49 (2009)
13. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, pp. 697–706. ACM, New York (2007)
14. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: Aberer, K., et al. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)

15. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems, I-Semantics 2011, pp. 1–8. ACM, New York (2011)
16. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Conference on EMNLP 2011, Edinburgh, Scotland, United Kingdom, pp. 782–792 (2011)
17. Leidner, J.L., Lieberman, M.D.: Detecting geographical references in the form of place names and associated spatial natural language. SIGSPATIAL Special 3(2), 5–11 (2011)
18. Tobin, R., Grover, C., Byrne, K., Reid, J., Walsh, J.: Evaluation of georeferencing. In: Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR 2010, pp. 7:1–7:8. ACM, New York (2010)
19. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002) (2002)
20. Bilhaut, F., Charnois, T., Enjalbert, P., Mathet, Y.: Geographic reference analysis for geographic document querying. In: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, HLT-NAACL-GEOREF 2003, Stroudsburg, PA, USA, vol. 1, pp. 55–62. Association for Computational Linguistics (2003)
21. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL 2005, Stroudsburg, PA, USA, pp. 363–370. Association for Computational Linguistics (2005)
22. Hatcher, E., Gospodnetic, O., McCandless, M.: Lucene in Action, 2nd revised edn. Manning (2010)
23. Waltinger, U., Breuing, A., Wachsmuth, I.: Interfacing virtual agents with collaborative knowledge: Open domain question answering using wikipedia-based topic models. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI 2011, Barcelona, Catalonia, Spain, July 16-22, pp. 1896–1902 (2011)
24. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: language-independent named entity recognition. In: Proceedings of the HLT-NAACL 2003, CONLL 2003, Stroudsburg, PA, USA, vol. 4, pp. 142–147. Association for Computational Linguistics (2003)
25. Ana Cristina Mendes, L.C., Lobo, P.V.: Named entity recognition in questions: Towards a golden collection. In: Calzolari, N. (ConferenceChair) Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the LREC 2010, Valletta, Malta. European Language Resources Association (ELRA) (May 2010)
26. Li, X., Roth, D.: Learning question classifiers: the role of semantic information. Nat. Lang. Eng. 12(3), 229–249 (2006)
27. Benefico, S.: Geo-related Information Extraction from natural language using YAGO. Technical report (2012)