Identifying and Extracting Named Entities from Wikipedia Database Using Entity Infoboxes

Muhidin Mohamed
School of Electrical, Electronic and Computer Eng.
University of Birmingham,
Edgbaston, Birmingham, United Kingdom

Mourad Oussalah
School of Electrical, Electronic and Computer Eng.
University of Birmingham
Edgbaston, Birmingham, United Kingdom

***Abstract*—An approach for named entity classification based on Wikipedia article infoboxes is described in this paper. It identifies the three fundamental named entity types, namely; Person, Location and Organization. An entity classification is accomplished by matching entity attributes extracted from the relevant entity article infobox against core entity attributes built from Wikipedia Infobox Templates. Experimental results showed that the classifier can achieve a high accuracy and F-measure scores of 97%. Based on this approach, a database of around 1.6 million 3-typed named entities is created from 20140203 Wikipedia dump. Experiments on CoNLL2003 shared task named entity recognition (NER) dataset disclosed the system's outstanding performance in comparison to three different state-of-the-art systems.**

Keywords—named entity identification; Wikipedia infobox; infobox templates; Named Entity Classification (NEC);

I. INTRODUCTION

The word named entity (NE) as used today in text mining and Natural Language Processing (NLP) was introduced in the Sixth Message Understanding Conference [1]. It represents a major part of all textual data covering proper names of persons, locations, organisations and corporate entities e.g, University of Birmingham, UK, Mount Everest, Mogadishu, David Beckham among others. Besides, Named entity classification (NEC) is the process of categorizing named entities to their corresponding classes (e.g. Person, Location, Organization). This is usually a supplementary step to the wider area of named entity recognition (NER). Although, NEs represent core components in natural language texts, they are still poorly covered in the state of the art language dictionaries. This might be due either to their ever-changing nature and dynamicity in which some named entities disappear while new ones emerge on regular basis, or to the fact that many NEs might be genuinely classified to more than one class, where one may encounter, for instance, several place names who are also person names, and/or corporate names. For example, if you search some of world's largest corporations such as Microsoft and Apple you may hardly find them in the state of the art knowledge networks such as WordNet. An improvement of named entity coverage are now being made in lexical semantic networks such as ConceptNet 5 [2]. More importantly, constantly updated live online repositories like Wikipedia [3] and Open Directory Project [4] do possess high named entity coverage than the aforementioned resources holding almost all object names. Therefore, in order to automatically handle NER or NEC tasks, the use of such repositories is inevitable. Challenges hindering an accurate NEC is not limited to their low coverage in the well-established language resources, but also include the ambiguity pervading the meaning of these entities [5], and entity linking [6], which have been subjected to intensive studies in recent years. This study is rather focused on improving NEC through addressing the coverage problem. To this end, current work advocates the use of Wikipedia utility for entity classification.

Strictly speaking, with the emergence of diverse natural language processing tools and the increasing need for automated text analysis, an important research has been conducted for the purpose of named entity classification in the past few years. In [7], authors used a bootstrapping method based on Wikipedia category to classify named entities containing Heidelberg Named Entity Resource (HeiNER) [8]. Nevertheless such classification might be undermined by the inconsistency of placing contributed articles by the authors in the most appropriate category. In a closely related study, Tkachenko et al. [9] carried out a fine grained classification for Wikipedia named entities. Though, their method correlates this study, they extracted many features for the classification including first paragraph of the article text, categories, template names, and other structured content tokens. This will demand a huge processing time when classifying large datasets. The closest work to ours is explored in [10] where researchers used structured information from infoboxes and category trees for the classification task. Despite this relatedness, their work differs from this study in terms of the overall classification methodology as well as the employed dataset where Portuguese Wikipedia was used in [8].

Finally, one shall also mention some seminal works on Wikipedia entity classification built on machine learning algorithms. Dakka et al. [11] used bag-of-words of Wikipedia articles with support vector machine (SVM) algorithm achieving a high F-score of (90%). Watanabe et al. [12] employed Conditional Random Fields to classify Japanese Wikipedia articles while Bhole et al. [13] combined heuristics with linear SVM for the same purpose. But the main drawback of machine learning related approaches lies in the requirement of a manually annotated training data, which is rather costly and complex task.

The main contribution of this paper consists in designing and testing a new simple named entity classification algorithm that only makes use of some structured information available in Wikipedia articles. Especially, unlike the aforementioned methods, the proposed NEC approach relies on the content

information of a single structured table, the infobox, but achieves a high score of accuracy and F-measure. The classification algorithm put forward in this study matches a predefined core entity attributes built from Wikipedia Infobox Templates (WIT) and entity specific attributes extracted from the related named entity Wikipedia article.

The rest of the paper is structured as follows. Section 2 covers Wikipedia structure and its containment of named entities. Section 3 copes with the proposed named entity classification approach using Wikipedia. Section 4 details the system experiments, highlighting the utilized dataset, results, and comparison with relevant state of the art systems. Finally, conclusions are drawn in Section 5.

II. WIKIPEDIA

A. Overview

Wikipedia is a freely available encyclopaedia with a collective intelligence contributed by the entire world community [14]. Since its foundation in 2001, the site has grown in both popularity and size. At the time of this study's experiment (April 2014), Wikipedia contains over 32 million articles [15] in 260 languages [16] where its English version has more than 4.5 million articles[1]. Its open collaborative contribution to the public arguably makes it the world's largest information repository.

Wikipedia contains 30 namespaces of which 14 are subject namespaces and two are virtual namespaces. Besides, each namespace has a corresponding talk namespace[2]. A namespace is a criterion often employed for classifying Wikipedia pages, using MediaWiki Software, as indicated in the page titles. Structurally, Wikipedia is organized in the form of interlinked pages. Depending on their information content, Wikipedia pages are loosely categorized as Named Entity Pages, Concept Pages, Category Pages, Meta Pages [8].

In recent years, there has been a growing research interest among the NLP and IR research communities for the use of this encyclopaedia as semantic lexical resources for tasks such as word semantic relatedness [17], word disambiguation [18], text classification [19], ontology construction [20], named entity recognition/classification [21], among others.

B. Named Entities in Wikipedia

Research has found that around 74% of Wikipedia pages describe about named entities [22], a clear indication of Wikipedia's high coverage for named entities. Each Wikipedia article associated with a named entity is identified with its name. Most Wikipedia articles on named entities offer useful unique properties starting with a brief informational text that describes the entity, followed by a list of subtitles which provide further information specific to that entity. For example, one may find information related to main activities, demography, and environment for Location named entities; education, career, personal life and so on for Person named entities. Relating concepts to that named entity are linked to the entity article by outgoing hyperlinks. Moreover, a semi-structured table, called infobox, summarizing essential attributes for that entity lives in the top right hand of each article [23]. It is the core attributes of the article infobox that this study stands on for the classification of named entities without any other prior knowledge. The snapshot in Figure 1 illustrates the Wikipedia article infobox related to "Google", which corresponds to a named entity of type Organization (http://en.wikipedia.org/wiki). The table summarizes very important unique properties of the entity in the form of attribute-value pairs. Consequently such tables are extracted, stored and analysed for the purpose of NE classification.

Fig. 1. Google Wikipedia article infobox[3]

III. THE CLASSIFIER

Using predefined core attributes extracted from Wikipedia Infobox Templates, a semi-supervised binary algorithm is developed. Being the main classifier, it predicts whether a particular named entity belongs to a given type. In other words, the classifier is designed to match named entities against these set of core class attributes (cf. Section A) and consequently identify these entities based on the outcomes of the matching process. The classification is achieved according to the following definition.

[1] http://en.wikipedia.org/wiki/Main_Page
[2] http://en.wikipedia.org/wiki/Wikipedia:Namespace
[3] http://en.wikipedia.org/wiki/Google

Definition: Let **ne** be a named entity in Wikipedia (**WP**) belonging to any of the three types, Person (**P**), Location (**L**) and Organization (**O**). If **XITA** denotes infobox template attributes[4] of type **X** and **IA(ne)** is the infobox attributes extracted from WP article associated with **ne**, then the classifier identifies **ne** type according to quantification (1).

$$T_{ne} = \begin{cases} P & if\ ne \in WP\ \&\ IA(ne) == PITA \\ L & if\ ne \in WP\ \&\ IA(ne) == LITA \\ O & if\ ne \in WP\ \&\ IA(ne) == OITA \end{cases} \quad (1)$$

Where T_{ne} stands for the type of named entity ne as identified by the classifier, while the operator "==" corresponds to array matching.

A. Defining Core Attributes

MeidaWiki team has developed infobox templates designed to guide contributing authors. The infobox templates contain the attribute labels to be filled by the authors with values when writing their Wikipedia articles on named entities. These attributes describe properties particular to each named entity type. For example, all location-based named entities should bear **coordinate** information. Similarly, infobox attributes for Person named entities include **birth date** and **place**. Table 1 lists a selected sample of these attributes for demonstration purpose. Essential attributes to each class, usually identified through manual investigation, are referred **Core Attributes**. The latter are used in the experiments to identify Wikipedia articles corresponding to named entities through matching the core attributes with the attributes extracted from entity infoboxes. Experimented core attributes are designated with stars in Table 1.

TABLE I. CORE ATTRIBUTES EXTRACTED FROM INFOBOX TEMPLATES

Person	Organization	Location
Birth_date*	Ceo, Founded*	Coordinates*
Birth_place*	Headquarters*	Population*
Spouce	Service_area*	Area*
Children	Industry , Profit*	Region
Relatives	Traded_as, revenue*	Country*
Occupation	Num_staff*,	timezone
Nationality	Num_employee*	iso_code
Parents	Established*	area_code
Education	Founder/chancellor*	settlement
Salary	{Post|under}graduates*	Leader_name
partner	{operating|net}income*	Leader_name

B. Accessing Wikipedia Database

To use Wikipedia as an external knowledge repository for named entity classification, a mechanism for accessing its database should be in place. Designed system's access to the encyclopaedia is summarized in Figure 2. Primarily there are two methods for accomplishing such data access; namely, either querying through web interface, or accessing a downloaded local Wikipedia dump.

For this study, query access method is used for the system evaluation. However, for the actual named entity extraction, a local access is made to a downloaded Wikipedia xml dump of February 2014. In implementing the query access method, this study partially adapts the Wikipedia Automated Interface [24] while the local access to the Wikipedia Dump is built on a MediaWiki dump Files Processing Tool [25]. The preference of query access over the local access for the evaluation is tied to the unsuitability of the dump files for random access as the dumps are primarily designed for sequential access.

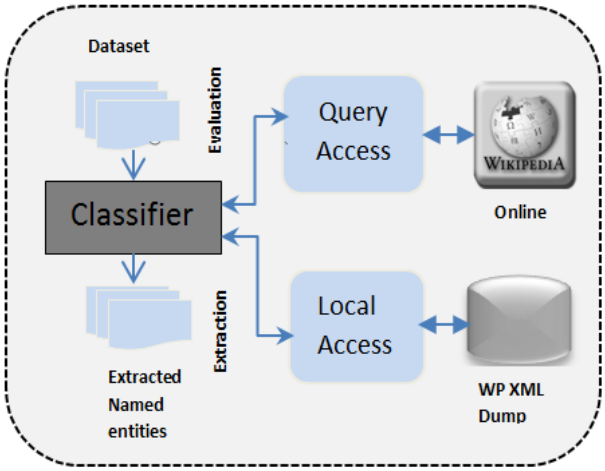

Fig. 2. Classifier's Access Mechanisms to Wikipedia

IV. EXPERIMENTAL SETUP

The proposed classifier system is implemented with Perl scripts in Linux environment. Entity core attributes derived from Wikipedia Infobox Templates represent the heart of the developed classification method. An illustration of the implementation scheme is given in Figure 4 (cf. the algorithm in Fig. 3). Each named entity has to go through three processing stages before it gets classified to its type. In stage 1, the Wikipedia article associated with that entity is retrieved while the extraction of its article's infobox forms stage two. At this stage, the scope of the processing text has been narrowed to the infobox. This semi-structured table is further parsed in stage 3 where tuples of attribute label-values are built from the infobox obtained in stage 2. Having organized the tuples in Perl Hashes, the matching process is now performed against the core attributes and the correct decision is made. The same process is repeated for every named entity to be identified. Figure 3 and Figure 4 better summarize the logical flow of the discussed classification methodology, in terms of pseudo-code and block diagram representation.

Algorithm1 ***WP Aided NE Classification***

```
1   ED← NE Evaluation Dataset
2   AV ← Infobox tempalte Attributes
3   C← {}
4   For all ( ne_i ∈ ED ) do
5     If ne_i ∈ WPDB then
6       A_nei ← RetreiveArticle(ne_i)
7       I_nei ← ExtractInfobox(A_nei)
8       For each v_j ∈ AV
9         If v_j =~ I_nei then
```

[4] These are the core attributes used for matching

```
10            cne ← ne_i#type(v_j)
11            Last;
12          Endif
13        Endfor
14      endif
15      C ← C υ { cne }
16    endfor
17    return C
```

Fig. 3. Perl-styled Pseudocode algorithm for Wikipedia infobox-based named entity classification.

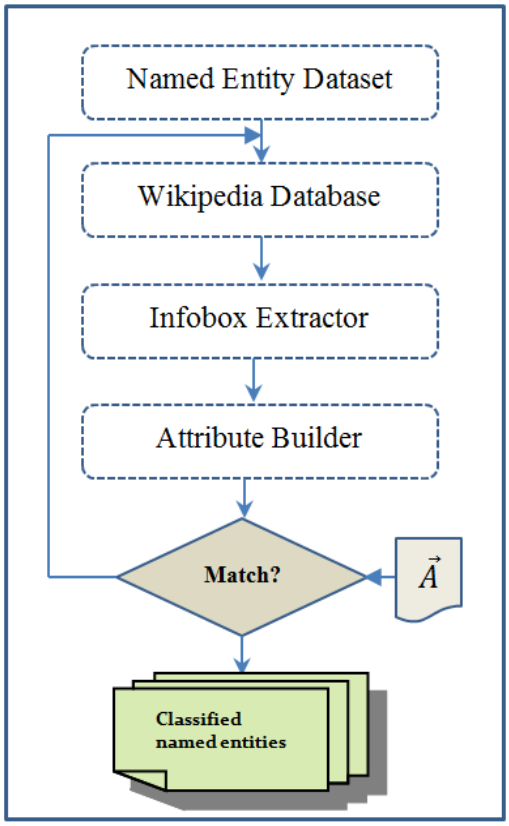

Fig. 4. Named entity Classifier Flowchart

A. Dataset

Experiments were conducted on two datasets. The first test-data comprises 3600 named entities with different proportions of the three considered entity types (PER, LOC, ORG), and was created from two data sources; namely, Forbes and GeoWordNet. Especially, all organization and person names were an excerpt of Forbes400 and Forbes2000 lists for richest American businessmen and world's leading public companies respectively[5]. On the other hand, Location named entities were sourced from GeoWordNet database. The second test-data uses CoNLL-2003 shared task named entity data[6]. The latter dataset, a standard publicly available dataset, has been selected for proper evaluation and comparison with state of the art techniques for Wikipedia NEC. Checking the coverage and the availability of all names with their surface forms in Wikipedia has been performed over all datasets prior to the experiments.

[5] http://www.forbes.com/lists/

[6] http://www.cnts.ua.ac.be/conll2003/ner/

B. Results and Discussion

The system tests were made in two rounds. In the first round the test dataset is divided into 4 smaller parts containing 100, 500, 1000, 2000 NEs all with different proportions of their types. This splitting has been performed for at least two reasons. First, this helps to securitize the data size effect on the observed parameters. Second, it reduces Wikipedia server's overhead with large data since all the testing and evaluation experiments used Query-based access to the online version of the encyclopaedia.

There are four possible outcomes that can result from the binary predictive classifier. In the first case, an entity that belongs to a type ***x*** might be classified as being of class ***x***, referred to as True Positive (TP). Secondly, A False Negative (FN) occurs when a named entity of type ***x*** is incorrectly identified as not falling in that type. Thirdly, there happens a case where a named entity does not belong to class ***x***, but classified as type x; a situation known as False Positive (FP). Lastly, when a non-member named entity of type ***x*** is correctly predicted as not falling in class **x**, it is referred to as True Negative (TN). Metrics for evaluating the classifier's performance will be based on the above mentioned outcomes.

Results of round 1 experiments are reported in Table 2, where the accuracy level is determined as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

TABLE II. RESULTS: ACCURACY WITH VARYING DATA SIZES

	Percentage Accuracy		
Dataset Size	Person	Location	Organization
100	96%	99%	97%
500	91.6%	95.4%	94%
1000	93.8%	94.2%	94.3%
2000	95.5%	93.9%	97.25%

The trend of the scores shown in Table 2 indicates that varying data sizes has little effect on the accuracy for the Person and Organization entity types. However a slight declination is observable in the case of Location names. Overall, round 1 experiments on test-data reveal that the classifier can achieve an average accuracy above 93% irrespective of the data size.

An examination of the misclassified proportion of the test data showed, that a number of factors contribute to the classifier's failure to identify some named entities. The most prominent factors were found to be the ambiguity of named entities and the absence of infoboxes from Wikipedia articles. Although there are machine learning like solutions to the ambiguity issue, using disambiguation like method, little can be done in the case of absence of infobox information. Possibly, the only sensible way to handle this matter is removing the underlying case (s) from the evaluation dataset in the validation and state of the art comparison stages.

TABLE III. ERROR CAUSING FACTORS AND SYSTEM MISIDENTIFICATIONS

Type	Ambiguity	No Infobox	Others
Person	46.7%	50%	3.3%
Location	11.1%	11.1%	**77.8%**
Organization	70.1%	20.8%	7.1

Disambiguation is the process of normalizing named entities that have multiple surface forms and identifying their referents. For instance, Birmingham may refer to the largest city in Alabama USA, or the second largest city in the United Kingdom. Error analysis related to system's misclassification highlighting factors leading to these errors is presented in Table 3. Through the error analysis, it is found that ambiguity in Organization and Person names extremely undermines the system performance. This is perhaps due to the use of abbreviations for larger organization names and the presence of common cultural names e.g John, Mohamed, shared by thousands of people in Wikipedia database. Disambiguating named entities in Wikipedia has been studied [5] and is still an active research problem.

Results of Table 3 have also shown the existence of a high proportion of Wikipedia named entities that lacked infoboxes information. Experimental results disclosed that **50%** of the unclassified Person entity articles are without infoboxes in Wikipedia. The figure is slightly lower for the other two considered entity types. As the system relies on information in the infobox, the absence of the infobox from any entity article makes the system unable to identify related named entity. Because of its importance, [26] proposed an author assistant tool for automatic suggestion of infoboxes for contributing authors.

In Table 3, the column designated by **Others** combines other factors including redirected pages, and technical difficulty of extracting the infobox due to the structure of some Wikipedia articles that lack regular patterns. Sometimes the availability of an infobox in an article does not guarantee the presence of the core attributes. The fact that some Wikipedia article infoboxes does not contain the core attributes such as coordinates made this factor to be the misclassification culprit for the largest percentage (77.8%) of unclassified Location named entities. This again precluded the classification of these entities on the basis of their core attributes.

Following the error analysis and prior to the second round of evaluative experiments, Wikipedia assisted disambiguation is used to exclude all ambiguous names. Similarly, all named entities whose Wikipedia articles lack infobox tables have been iteratively removed from the evaluation dataset.

In the second round, experiments were conducted using named entities constructed from CoNLL-2003 shared task data for named entity recognition to observe three of the traditional information retrieval metrics namely; precision, recall, and F-measure. Precision is the proportion of classified named entities that belong to the target type. It is defined by the relationship in expression 3.

$$P = \frac{TP}{TP + FP} \qquad (3)$$

Likewise, recall (exp. 4) measures the proportion of named entities of a given type which has been correctly classified.

$$R = \frac{TP}{TP + FN} \qquad (4)$$

Due to the trade-off between precision and recall, an F-measure has been developed as proper measure that combines the effect of the metrics as formulated in equation 5.

$$F = \frac{2RP}{P + R} \qquad (5)$$

The overall classifier results in terms of these three metrics are summarized in Table 4. The F-measure scores of locations and organizations indicate that the selected core attributes represent good classification criteria for identifying Wikipedia entities. Again, this study's results confirmed that these attributes are mainly added by article contributors when authoring Wikipedia articles through adapting infobox templates. Person names achieved the highest F-score as ambiguity of these has been accounted for.

TABLE IV. OVERALL CLASSIFIER RESULTS

Type	Precision	Recall	F-score
Person	1	0.98	0.99
Location	0.99	0.95	0.97
Organization	0.94	0.97	0.96

C. State-of-the-Art Comparison

Comparing the study's infobox based matching approach with related state of the schemes for named entity classification and extraction is not trivial. Major discrepancies arise from the peculiarity of each approach in terms of the Wikipedia features (article text, links, categories, infoboxes) used for the entity identification. In addition, there might be significant differences in the evaluation data and Wikipedia language in the event of language dependent schemes. Nevertheless, a rough approximate comparison of the system with three baselines is provided in Table 5. The criteria for choosing these baselines are their closeness to the system in terms of their use of infobox information and related features.

TABLE V. COMPARING F-SCORES WITH BASELINE SYSTEMS

System	PER	LOC	ORG
Bhole [13]	72.7	70.5	41.6
Gamallo[10]	88	63	73
Tardif [27]	95	99	93
This system	99	97	96

Table 5 compares the outcomes of the overall classification system in terms of F-score for each type of named entity to three state of art classification approaches (baselines). The baselines use infobox data as one of their classification features; whereas this system is entirely built on infobox attribute matching. Despite that, it is evident that it outperforms all baselines except [27] where a high F-score is reported for location based named entities. However, there is still a room for improvement to extend the work in identifying *Miscellaneous* named entities and further subcategorizing the

main entity types to subcategories which have been considered by many state of the art systems.

D. NE Extraction from Wikipedia

If any named entity with an entry in Wikipedia can be identified, then hypothesis on the likelihood of recognizing all Wikipedia articles on these entities can be reached. Therefore, the proposed classification algorithm is applied on the English Wikipedia dump dated third February 2014. Table 6 shows the number of each named entity type extracted from Wikipedia database. The number of named entities obtained through this approach (1575966) significantly outnumbers the figure of Wikipedia articles on named entities (1547586) derived from the same database in [8]. One may argue that this has been an earlier study while Wikipedia is constantly growing in size. This is true to an extent, however this study has only considered three types of named entities while [8] contains Miscellaneous named entities in addition to the three considered by this work. The generated database of named entities can be used as a training data for supervised classification strategies.

TABLE VI. SUMMARY OF EXTRACTED WIKIPEDIA NES

Person	Location	Organization	total
620790	290134	665042	1575966

V. CONCLUSION

A Wikipedia-based approach for predicting three types of named entities namely; Person, Location and Organization using article infoboxes is presented. Unlike common state of the art approaches which rather employ a set of multiple features such as article text, categories, links, among others, this study relies on a single feature consisting of the structured information in the infobox table. This has significantly reduced the classifier's processing time, which would be useful for delay sensitive applications requiring identification of designated names. Despite the use of a single feature, the proposed approach achieves a classification accuracy of above 97% with 3600 named entities and CoNLL-2003 shared task NER dataset used to validate the classifier's performance. Applying the same algorithm on Wikipedia database has resulted in the extraction of around 1.6 million named entities belonging to these three types. As a future work, the ongoing study aims to extend the infobox-based entity identification to generate a fine-grained entity classes in which each of the main types can be further subdivided into multiple subtypes.

REFERENCES

[1] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A Brief History," in COLING, 1996, pp. 466-471.

[2] H. Liu and P. Singh, "ConceptNet—a practical commonsense reasoning tool-kit," BT technology journal, vol. 22, pp. 211-226, 2004.

[3] F. Wu and D. S. Weld, "Open information extraction using Wikipedia," in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 118-127.

[4] J. Liu and L. Birnbaum, "Measuring semantic similarity between named entities by searching the web directory," in Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence, 2007, pp. 461-465.

[5] S. Cucerzan, "Large-Scale Named Entity Disambiguation Based on Wikipedia Data," in EMNLP-CoNLL, 2007, pp. 708-716.

[6] W. Zhang, J. Su, C. L. Tan, and W. T. Wang, "Entity linking leveraging: automatically generated annotation," in Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 1290-1298.

[7] J. Knopp, "Extending a multilingual Lexical Resource by bootstrapping Named Entity Classification using Wikipedia's Category System," Proceedings of the Fifth International Workshop On Cross Lingual Information Access, vol. 5, 2011, 35-43.

[8] W. Wentland, J. Knopp, C. Silberer, and M. Hartung, "Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration," in Proceedings of the Sixth International Conference on Language Resources and Evaluation , LREC, Marrakech, 2008,

[9] M. Tkachenko, A. Ulanov, and A. Simanovsky, "Fine Grained Classification of Named Entities In Wikipedia," HP Laboratories Technical Report-HPL-2010-166, 2010.

[10] P. Gamallo and M. Garcia, "A resource-based method for named entity extraction and classification," in Progress in Artificial Intelligence, ed: Springer, 2011, pp. 610-623.

[11] W. Dakka and S. Cucerzan, "Augmenting Wikipedia with Named Entity Tags," in IJCNLP, 2008, pp. 545-552.

[12] Y. Watanabe, M. Asahara, and Y. Matsumoto, "A Graph-Based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields," in EMNLP-CoNLL, 2007, pp. 649-657.

[13] A. Bhole, B. Fortuna, M. Grobelnik, and D. Mladenić, "Extracting Named Entities and Relating Them over Time Based on Wikipedia," Informatica, vol. 31, 2007, 463-468.

[14] T. Zesch, I. Gurevych, and M. Mühlhäuser, "Analyzing and accessing Wikipedia as a lexical semantic resource," Data Structures for Linguistic Resources and Applications, pp. 197-205, 2007.

[15] WikiMedia. (2014). Wikipedia Statistics. Available: https://stats.wikimedia.org/EN/Sitemap.htm#comparisons

[16] P. Shachaf and N. Hara, "Beyond vandalism: Wikipedia trolls," Journal of Information Science, vol. 36, pp. 357-370, 2010.

[17] M. A. Hadj Taieb, M. Ben Aouicha, and A. Ben Hamadou, "Computing semantic relatedness using Wikipedia features," Knowledge-Based Systems, vol. 50, pp. 260-278, 2013.

[18] A. Bawakid, M. Oussalah, N. Afzal, S.-O. Shim, and S. Ahsan, "Disambiguating Words Senses with the Aid of Wikipedia," Life Science Journal, vol. 10, 2013, 1414-1426.

[19] P. Wang, J. Hu, H.-J. Zeng, and Z. Chen, "Using Wikipedia knowledge to improve text classification," Knowledge and Information Systems, vol. 19, pp. 265-281, 2009.

[20] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," Artificial Intelligence, vol. 194, pp. 28-61, 2013.

[21] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from Wikipedia," Artificial Intelligence, vol. 194, pp. 151-175, 2013.

[22] J. Nothman, J. R. Curran, and T. Murphy, "Transforming Wikipedia into named entity training data," in Proceedings of the Australian Language Technology Workshop, 2008, pp. 124-132.

[23] D. Lange, C. Böhm, and F. Naumann, "Extracting structured information from Wikipedia articles to populate infoboxes," in Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 1661-1664.

[24] B. C. Ed Summers. (2011). WWW::Wikipedia - Automated interface to the Wikipedia. Available: http://search.cpan.org/~bricas/WWW-Wikipedia-2.01/

[25] T. Riddle, "Parse::MediaWikiDump- Tools to process MediaWiki dump files," 2010.

[26] A. Sultana, Q. M. Hasan, A. K. Biswas, S. Das, H. Rahman, C. Ding, and C. Li, "Infobox suggestion for Wikipedia entities," in Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 2307-2310.

[27] S. Tardif, J. R. Curran, and T. Murphy, "Improved text categorisation for Wikipedia named entities," in Australasian Language Technology Association Workshop 2009, 2009, p. 104-109.