DE GRUYTER OPEN

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **18**, No 1
Sofia • 2018 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2018-0009

An Automatically Generated Annotated Corpus for Albanian Named Entity Recognition

Klesti Hoxha, Artur Baxhaku

University of Tirana, Faculty of Natural Sciences, 1001 Tirana, Albania
E-mails: klesti.hoxha@fshn.edu.al baxhaku@fshn.edu.al

***Abstract*:** *Named Entity Recognition* (*NER*) *is an important task in many NLP pipelines. It has become especially important for knowledge bases that power many of the nowadays information retrieval systems. In order to cope with the high demand for annotated training corpora for supervised NER systems, automatic generation approaches have been proposed. In this paper we report on the first automatically generated NE annotated corpus for Albanian. News articles from Albanian news media were used as a document source. They were automatically tagged using a custom generated gazetteer from the Albanian Wikipedia. Our evaluation results show that this corpus can be used as a baseline corpus for human annotated ones or as a training corpus where no other is available.*

***Keywords*:** *Named entity recognition, natural language processing, language corpora, semi-automatic annotation, information extraction.*

1. Introduction

Named Entity Recognition (NER) is the task of identifying named entities (people, locations, organizations, etc.) in text documents. A lot of work has been done in this regard in the past years [1], introducing various approaches of performing this task. Most of these approaches are machine learning based and make use of large annotated corpora of named entity mentions [2, 14] for training their models. The creation of these corpora is a costly and error prone approach that includes separate quality assurance activities.

NER has become a very important preliminary step when dealing with entity linking, the task of identifying entities in a knowledge base that are mentioned in a text document [4]. Many information retrieval systems nowadays make use of these knowledge bases for improving the offered user experience [5]. Therefore, the availability of NER tools strongly affects the inclusion of knowledge bases data in these systems.

While the state-of-the-art NER approaches have achieved a very high accuracy when trained with in-domain corpora, the general availability of these corpora is

limited to the most popular languages. In order to cope with the high demand of such NER tools for each language and also reduce the cost of human annotated corpora creation, semi-automated annotation methods have been proposed. Many of them rely on WikiPedia (WP) as a collaboratively written body of documents that mention plenty of Named Entities (NE) [4, 5].

Albanian is not a well resourced language in terms of Natural Language Processing (NLP). Many previous works in this regard were mostly experimental and there are few generally available NLP resources that deal with the Albanian language. There are few open source tools that support Albanian and to the best of our knowledge there is no high quality publicly available annotated corpus for Albanian NER. Nevertheless, it has been shown that the most common NER approaches work well for Albanian when trained appropriately [3, 9], but the annotated corpora that were used in these works have a modest size.

In this paper we describe our approach of generating automatically annotated NER corpora (silver corpus) for the Albanian language. To the best of our knowledge this is the first large enough Albanian NER annotated corpus. We use the Albanian version of Wikipedia for generating a custom gazetteer and automatically annotate a collection of news articles published in Albanian online news media using it. The generated corpus has been evaluated using Apache OpenNLP (**https://opennlp.apache.org/**), an open source natural language processing toolkit that uses a maximum entropy model [2] for named entity recognition. The same tool has also been successfully used in a related work [8] of Albanian NER, but using a human annotated corpus of historical-political domain documents.

We used the three main entity categories described in the ConLL NER Task [1]: people, locations, organizations. The decision if an entity belongs to a specific category is based on the ConLL annotation guideline (**http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt**). This approach is in the same line with many other related works. The collection of news articles that was used as a text corpus covers a variety of topics (domains).

An automatically generated NE corpus cannot be considered a gold standard one; however it can serve as a basis for crowd sourcing named entity tagging activities, or can be used as a training corpus if there is no other one available. In this case, the achieved recall is usually lower in comparison with NER models trained with gold standard corpora (human annotated).

In the following sections, after giving an overview of related works, we describe the NER gazetteer generation method. Then the corpus and its generation methodology are described. The paper is concluded with the experimental evaluation of the generated corpus.

2. Related works

There have been many previous works that report on semi-automatically generated NE corpora. Even though the quality of these kinds of corpora is usually lower in comparison with human annotated ones, they have a clear advantage in terms of

updatability. These approaches make use of custom implemented tools for extracting named entity gazetteers from semi structured documents, or directly annotating them by using name entity detection heuristics. These tools can be reused in text document collections that are updated and enlarged frequently (i.e., Wikipedia), producing so an updated version of the NE annotated corpora. It is also possible to annotate texts from various domains using automatically generated gazetteers; however the recall in this case may be lower because many named entities are domain specific.

Gazetteers in a NER context are lists of names, locations, organizations, etc. They can be used for automatically annotating texts with mentions of the entries of a gazetteer in question. Toral and Munoz [6] proposed a methodology for automatically generating gazetteers from Wikipedia, using individual articles (pages) as candidate entities. They make use of Part-of-Speech (PoS) taggers and WordNet for identifying entities based on language clues present in the first sentences of each article. The generated gazetteer was evaluated using a manually annotated list of articles. Table 1 shows the average achieved precision and recall (for each NE category) in their best performing experiment. Even though the proposed approach is language independent, it actually depends on the availability of NLP tools for the language in question (PoS taggers and WordNet).

Table 1. Toral and Munoz [6] NE gazetteer evaluation results

NE class	Precision	Recall
Person	74%	54%
Location	77%	61%
Organization	48%	11%
Average	**66%**	**42%**

Richman and Schone [9] describe an approach of generating NER training data from Wikipedia without making use of PoS taggers or WordNet. They focus instead on links to other WP pages. The observed links were: categories, redirect pages, disambiguation pages, links to the same WP article in another language, or links to other articles in the same language. The entity type categorization is based mostly on the category links present in each article. When dealing with non English articles, the categorization is based on the English version of that article (if a link to that is available) or the English version of the linked categories. Other than people, locations, and organizations, their corpus contains also other entities like dates or money. The evaluation was done using the generated annotated corpus to train a NER tool for various languages. The performed experiments achieved an F-Score comparable with the one achieved by using human annotated training corpora.

Nemeskey and Simon [7] report on an automatically generated NE corpus for English and Hungarian. Similarly to the work of Richman and Schone [9], they use links to other Wikipedia articles for identifying entities. In order to categorize entities, in addition to WP categories, they also make use of DBPedia [10], a knowledge base of entities mentioned in Wikipedia. After automatically generating a list of entities from WP articles, they annotate named entities in Wikipedia articles using the article link structure (i.e., the text of a link

that points to a country is annotated as a location entity). Most of the categorization method is based on DBPedia ontology classes mapped manually to WP categories. Because of this, when dealing with the Hungarian Wikipedia, they had to link individual WP articles to their equivalent in English. This was done because there is no Hungarian version of DBPedia. The annotated corpus contains the three most used NE categories (person, location, and organization) and the MISC category as defined in ConLL [1]. The generated corpus was evaluated using a maximum entropy entity tagger. The achieved precision and recall for Hungarian was 91% and 89% respectively when evaluating against the same Hungarian WP dataset. When using a human annotated Hungarian dataset of news articles, the achieved precision and recall was 63% and 70% respectively. The authors pointed out that a decrease in precision and recall is normal when classifying documents with a model trained from a corpus out of their domain.

Automatically generated NE corpora have also been tried in highly specialized domains like medical documents or scientific articles. In [11] is described an approach of automatically generating a NE training corpus of medical entities for Italian written clinical records. The authors used two different methods for generating the corpus: machine translation from English annotated corpora and directly labeling entities in Italian texts (clinical records) using a dictionary based approach. When using a dictionary based approach they used approximate matches (*n*-gram or lemma based). The achieved precision and recall for the best experimental setting was 80% and 67% respectively. When no gazetteer/dictionary is available and it is not possible to automatically generate one, it is possible to use a rule-based approach (i.e., using regular expressions) to automatically annotate a corpus. Such an approach has been reported in [12] for automatically annotating a medical corpus with physiological measurements.

When dealing with scientific documents, other named entity categories arise, i.e., scientific concepts. Prokofyev, Demartini and Cudré-Mauroux [13] describe a method of recognizing such entities in physics and computer science scientific articles. They make use of PoS tags and *n*-gram based features for extracting candidate entities. Dealing with highly specific entity categories, their task is actually similar to a "key terms extraction" one. The actual named entity extraction is held by a trained classifier that selects them from a list of candidate entities. Therefore, this method can be considered as a combination of an unsupervised and supervised approach. The achieved results showed that this approach noticeably outperforms maximum entropy based NER ones.

To the best of our knowledge, there are no previous works that report on an automatically generated NE annotated corpus for the Albanian language. This also holds for NE gazetteers in Albanian. We also couldn't identify any previous work that automatically extracts information from the Albanian version of Wikipedia. Previous NER works on Albanian documents used human annotated NE corpora of a limited size.

In [3] a NER attempt is described for Albanian news articles using a human annotated corpus of modest size (about 1,000 sentences). It was created by non expert annotators using an *n*-gram based annotation strategy. The experiments were

conducted using Stanford NER (**https://nlp.stanford.edu/software/CRF-NER.html**), a CRF [14] based named entity recognizer, achieving an average F-Score of 70%.

A similar work on Albanian NER was reported by S k ë n d u l i and B i b a [8]. They also created a human annotated NE corpus of modest size in the historical/political domain (about 3,000 sentences). Experiments were conducted using Apache OpenNLP maximum entropy based classifier. The average achieved F-Score was 75%.

3. Albanian named entity gazetteer generation

In order to automatically annotate our corpus, we generated a gazetteer of people, locations, and organizations from the Albanian Wikipedia. We used the definitions available in the ConLL NER task [1] for the three NE categories in question. Albanian is not a well resourced language in terms of natural language processing. To the best of our knowledge, there is no publicly available PoS tagger and the Albanian version of WordNet (**https://fjalnet.com/**) is very limited. For this reason, we did not follow an approach that makes use of robust NLP resources, but focused instead in the metadata available in the WP dumps (**https://dumps.wikimedia.org/sqwiki/20170601/**).

Our approach was similar to the ones described in [7] and [9]. However, we have followed a slightly different approach in order to deal with the current structure of the WP dumps and also used Wikidata [15] as a knowledge base.

We considered individual WP article titles as candidate entities. At the current time of writing there are over 70,000 articles and over 20,000 redirect pages to these articles in the Albanian Wikipedia. We used the following resources of the WP dumps for generating our gazetteer:

Article dumps offered in XML format. These were parsed using the WikiExtractor (**https://github.com/attardi/wikiextractor**) Python tool. We stripped all formatting tags except “intrawiki links” to other pages in Albanian WikiPedia.

The page table (**https://www.mediawiki.org/wiki/Manual:Page_table**) of the WP SQL dumps. It contains an index of all WP pages available in the XML dumps. We focused only on article pages (including redirects/aliases of them).

The categorylinks table (**https://www.mediawiki.org/wiki/Manual: Categorylinks_table**) of the WP SQL dumps. It contains all categories that a page belongs to.

The Langlinks table (**https://www.mediawiki.org/wiki/Manual: Langlinks_table**) of the WP SQL dumps. It contains all “interwiki links”, links that lead to the same article in other languages. We use this table for getting the English version of an article if it is available.

When generating the gazetteer, we did not deal with regular polysemy. Using the WP article titles as candidate entities, we assumed that each one of them that is a named entity belongs only to one NE category. This limitation might lower the

recall for individual NE categories, but as reported in other works, this is usually the case with automatically generated NE corpora.

Similarly to [9], we used page categories as a first step on categorizing the entities in our three focus categories (person, location, organization). This was based on the category links found in the *categorylinks* table described above. We manually compiled a list of categories that indicate the belonging to each of the NE categories in question. This process was also aided by a blacklist (WP articles may belong to multiple categories), in order to avoid false NE category labeling. For example, titles of articles belonging to the politician or writer categories (in Albanian) were tagged as person, while titles of articles belonging to city or village categories were tagged as location. In the blacklist categories we included the ones that do not indicate person, location, or organization entities, i.e. flags, food, novel, books, movies, etc.

All candidate entities (WP article page titles) that were left untagged from the first step went through an annotation procedure that uses Wikidata ontology for classifying the entities. This is similar to the work of Nemeskey and Simon [7], but they used DBPedia instead. Wikidata, launched in 2012, in comparison to DBPedia is collaboratively edited in a similar way that Wikipedia itself is updated. So instead of extracting facts from WP, Wikidata items usually feed WP articles with facts or claims. Each claim (property) in Wikidata can also have a list of references.

An advantage of using Wikidata is that is naturally interlinked with WP. Wikipedia articles usually contain a link to the corresponding Wikidata item. We used these links in order to run our Wikidata based categorization step. When a link was missing, we checked the English version of the article in question for a link to the corresponding Wikidata item. Wikidata items are offered in various languages, but using the same URI. Concretely, we used the *instance of* (P31) and *subclass of* (P279) properties of the Wikidata items in order to identify person, location, or organization entities. The subclass hierarchy was followed until a relevant class was encountered (we limited the depth of this search to three levels).

Table 2 details the Wikidata classes (name and unique identifier) that were used for identifying each NE category. This mapping was manually defined by us. At the end of this step, all Albanian WP articles titles that were left untagged are labeled as UNKnown (UNK) and were not included in the generated gazetteer.

As a last step, all Albanian WP redirect pages are labeled with the same tag as the page they redirect to. This extends the gazetteer with a list of aliases for many of the included entities. Finally, the gazetteer is extracted. We have stored it as a serialized Python dictionary that will be used in the annotation step.

Fig. 1 gives an overview of our gazetteer generation approach. The WP categories/Wikidata classes mappings to NE categories was fine tuned using a quality check that consisted of a manual inspection of the generated gazetteer. Table 3 displays the total number of entities in the generated gazetteer for each category and the respective densities (calculated as the percentage in relation to the total number of NE candidates). More than half of the article titles of the dataset in use (WP dumps) resulted to belong to the three NE categories in question. The

organization category contains the lesser amount of entities, while about 40% of the WP article titles belong to the location class.

Table 2. Mapping of Wikidata classes to the corresponding NE category

Wikidata classes	NE category
Human (Q5), person (Q215627)	Person
Country (Q6256), district (Q149621), island nation (Q112099), capital (Q5119), city (Q515), municipal unit (Q28017630), administrative region (Q3455524), geographic location (Q2221906), location (Q17334923), mountain (Q8502), terrain (Q14524493), geographical object (Q618123), river (Q4022), watercourse (Q355304), land waters (Q863944), island (Q23442), landform (Q271669), landscape (Q107425), territorial entity (Q1496967), beach (Q40080), coastal landform (Q19817101), continent (Q5107), national park (Q46169), fortress (Q57831), fortification (Q57821), military building (Q6852233), church (Q16970), temple (Q44539), teaching hospital (Q1059324), hospital (Q16917), public building (Q294422), sports venue (Q1076486), football stadium (Q1154710), stadium (Q483110)	Location
Political party (Q7278), political organization Q7210356), organization (Q43229), international organization (Q484652), ministry (Q192350), rock band (Q5741069), musical ensemble (Q2088357), newspaper (Q11032), magazine (Q41298), school (Q3914), general education school (Q12379547), commercial building (Q655686), shopping mall (Q11315), university (Q3918), higher education institution (Q38723), association football club (Q476028), football club (Q17270000), sports club (Q847017), sports organization (Q4438121), court (Q41487), bank (Q22687), administrative territorial entity (Q56061), government agency (Q327333), regional organization (Q4120211), broadcaster (Q15265344)	Organization

Table 3. Gazetteer NE categories count

NE class	Count	Density
Person	13,639	14.9%
Location	36,368	39.8%
Organization	2,903	3.2%
Total	52,910	57.9%

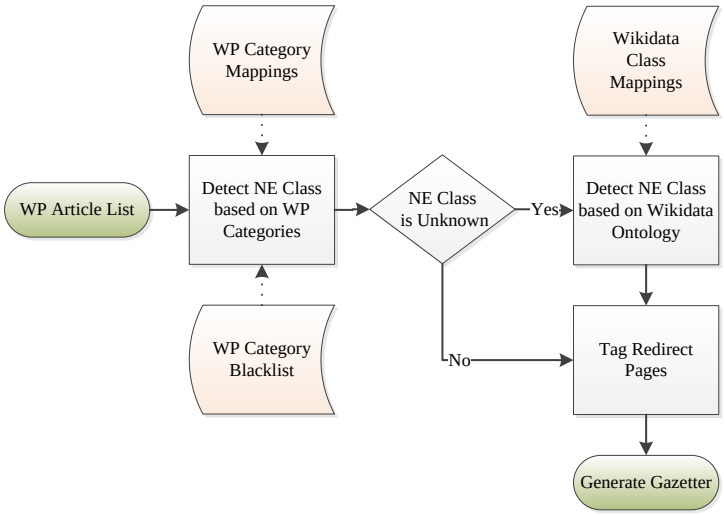

Fig. 1. NE gazetteer generation

4. Annotated corpus description

We used news articles published in Albanian online news media as a document source for our corpus. The aim is to achieve a high precision and recall for news data sets or news retrieval use cases. In order to improve the recall of the algorithm trained with the generated corpus we use the following resources together with the generated gazetteer:

List of most common Albanian first names (366 entries).

List of most common Albanian last names (358 entries).

List of phrases in Albanian that indicate a person entity (i.e., Mr., Ms., Dr., Professor).

List of phrases in Albanian that indicate a location entity (i.e., county, neighborhood, city).

List of phrases in Albanian that indicate an organization entity (i.e., university, department, institute).

Lacking a PoS tagger for Albanian, we limited the candidate entities extraction from text, to a capitalized text sequence (excluding articles) approach. Because of this, we noticed problems with the annotation of words at a beginning of the sentence. The way to deal with these cases is usually PoS tagging based [7], so we leave this for future works. Lacking a PoS tagger, also made difficult dealing with inflected words. In order to cope with this, we tried matching candidate entities (capitalized sequences of words) with the gazetteer entries using an approximate match (inflection in Albanian is usually done through word endings).

We generated three different annotated corpora, one for each of the NE categories. This is recommended by the Apache OpenNLP NER toolkit, the one that was used for the evaluation of our corpus. We also followed the annotation format of the same toolkit. It has one sentence per line and named entities are annotated using annotation tags. A blank line marks the beginning of a new document. An excerpt of the annotated corpus is shown in Fig. 2.

<START:person> Elseid Hysaj <END> ka shprehur vlerësimin e tij për ish mbrojtësin dhe kapitenin legjendar të Interit , <START:person> Javier Zanetti <END>.
Si futbollist , kam pëlqyer gjithmonë <START:person> Javier Zanettin <END>, si një lojtar dhe profesionist i madh.
“ Shpresoj të arrij në nivelet e tij “, ka thënë <START:person> Hysaj <END>.
...
Një grup studentësh nga <START:location> Suedia <END> po studiojnë ndryshimet që ka pësuar së fundi <START:location> Gjirokastra <END> për punimet e tyre shkencore .
Laura, me origjinë nga <START:location> Kosova <END>, thotë se ky është momenti kur duhet të mbrohen elementët tradicionalë të banesave.
“ Duhet të ruhen shtëpitë e vjetra që të ruhet autenticiteti ”, thotë Laura .

Fig. 2. Annotated corpus excerpt

For word and sentence tokenization we used the *polyglot toolkit* (**http://polyglot.readthedocs.io/**) implementation of the Unicode text segmentation algorithm [16]. We noticed that the tokenization produced by this algorithm is not perfect for Albanian, however there are no available corpora for training a machine learning based algorithm for tokenization.

Fig. 3 gives an overview of the used annotation process. The candidate entities consisted of sequences of text with all words capitalized (except articles). For each of them we applied the following steps.

Firstly, we tried to find an exact match in the gazetteer. If this did not succeed, next step is to find an approximate match in the gazetteer in order to deal with grammatical cases or other word forms (inflection). For this purpose the *difflib* (**https://docs.python.org/2/library/difflib.html**) Python library was used. It implements Ratcliff-Obershelp similarity algorithm [17]. We used a 0.75 cut-off similarity threshold. The first two steps are applied to all named entity categories. The next steps differ based on the NE category in question.

For the organization category, we label as an organization all candidate entities at least 2 characters long, that have all characters uppercased. This was done because news articles contain a lot of acronyms. A blacklist was used to remove common uppercase strings that are not entities. The last step for the organization category is to look for organization indicator strings in the candidate entities. If there is a match, they are labeled as organizations.

For the person category, as a next step, we try to find occurrences of person indicator phrases. As a last step we try to match the candidate entity with a combination of the list of common first names and last names. This is done with an approximate matching using a 0.8 similarity cut-off.

For the location category, as a last step, we try to match the candidate entities with the list of location indicators.

We keep track of the identified named entities and use this list as a last try to match any of the candidate named entities that were not tagged from the previous steps. This is done because very often in news articles, named entities are mentioned again in a non full form. For example, a person may be mentioned only by her first or last name.

We do not discard sentences that do not contain annotated entities. This is done in order to keep the original word distribution in the training corpus. However, a document with less than two sentences that contain annotated entities is discarded. Table 4 shows an overview of the generated corpus contents.

Table 4. Generated corpus contents overview

Sentences	Person	Location	Organization
Sentences	123,395	102,197	123,898
Annotated Sentences	25,007	25,008	25,006
Total Entities	30,235	33,945	31,898
Entities/Sentence	0.25	0.33	0.26
Size (MB)	15.9	19.8	18.9

5. Evaluation

Our generated corpus was evaluated using Apache OpenNLP NER toolkit. It implements a supervised machine learning algorithm for NER that is Maximum Entropy (ME) based. The maximum entropy based NER methods have achieved state-of-the-art results for many other languages and even some first attempts in Albanian [2, 7, 8].

ME models make use of various features for their classification approach. Examples include dictionary based features (compare candidate entities with a vocabulary), word based features (i.e., check if a word is capitalized), transition features (check prior NE tags) [6]. For our experiments we used the default feature generators used by the OpenNLP NE finder.

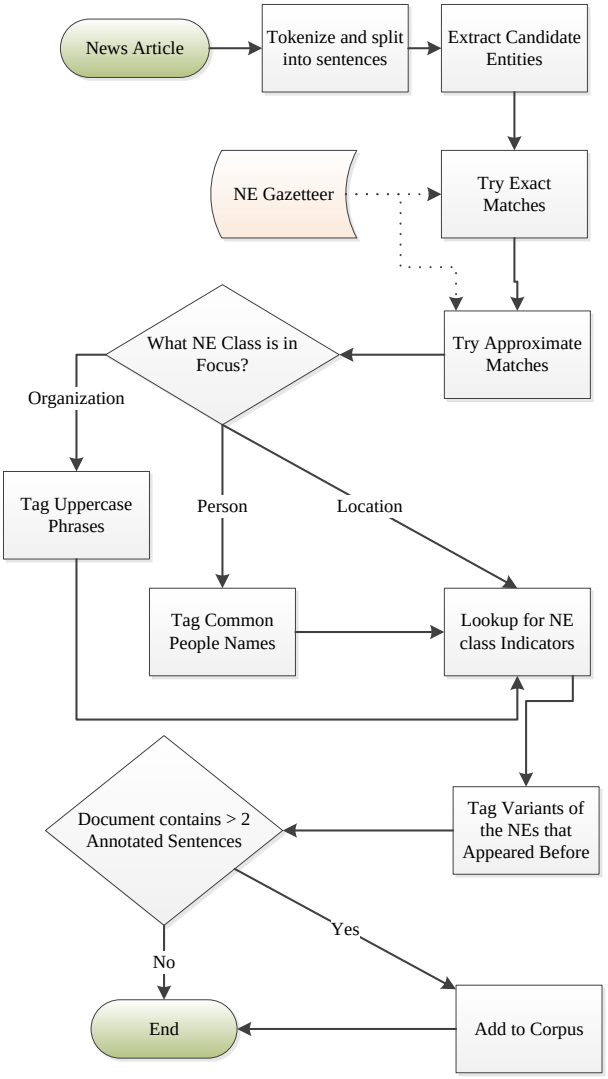

Fig. 3. News article annotation process

They are mostly context-based, considering the surrounding window of a token, previously occurred outcomes of a word, distribution of outcomes, bigram based features, and features considering the position of a word in a sentence.

Based on preliminary testing we compiled a blacklist of entities that were wrongly labeled as named entities. This reduced the number of false positives that resulted by our automatic annotation approach. The blacklist included individual words/phrases and restrictions on the character length of the named entities depending on the NE category in question.

We firstly evaluated our corpus against itself using a 10-fold cross validation procedure [18] that splits the dataset in a 9/1 training/testing setting. This was done using the built-in cross validator provided by Apache OpenNLP. We experimented with various configuration settings of the gazetteer generation and automatic annotation procedure. The best achieved results are displayed in Table 5. They are better than the ones achieved with human annotated corpora reported in [3] and [8]. The achieved precision is higher in comparison with experiments that used automatically annotated corpora for other languages [7, 9], while the recall is slightly lower.

This lower recall may have resulted because of the limitations made to our approach due to the lack of proper NLP tools for Albanian (PoS tagger, tokenizer, WordNet). Not dealing with regular polysemy may have also lowered the recall for the individual NE categories in question. Another possible explanation of the slightly lower achieved recall is the fact that we evaluated against a news dataset annotated using a NE gazetteer extracted from Wikipedia. Many named entities mentioned in news articles are not yet present in the Albanian Wikipedia. Performing more than one iteration for generating the annotated corpus (using the first one for creating the initial training corpus), might also have improved our results.

Table 5. 10-fold cross validation results

Data	Precision	Recall	F-Score
Person	89.03%	72.51%	79.92%
Location	91.92%	77.00%	83.80%
Organization	93.92%	75.03%	83.42%
Average	**91.62%**	**74.85%**	**82.38%**

We also evaluated our corpus using a human annotated corpus of 1000 sentences of news articles as a testing dataset [3]. It needs to be noted that this corpus has not been created by expert annotators, but by voluntary computer science students instead. Furthermore, no quality control mechanism has been applied to the annotation procedure of it. Experiments were conducted using the built-in evaluation tool provided by Apache OpenNLP (the testing data need to be in the same format as the training data). Again, we experimented with various configuration settings. The best achieved results are displayed in Table 6. As expected, they are lower in comparison with human annotated corpora training approaches, however are good enough for use cases that use the automatically annotated corpus as a baseline.

Table 6. Evaluation results against human annotated test data

Data	Precision	Recall	F-Score
Person	74.70%	24.75%	37.18%
Location	87.79%	51.85%	65.20%
Organization	76.04%	45.06%	56.59%
Average	**79.51%**	**40.55%**	**52.99%**

The recall for the person category is very low. Through manual inspection we noticed that the main cause of this could be the candidate NE extraction methodology that we used (sequence of capitalized words excluding articles). For example, there are cases when a location name is written next to a person name and they are both capitalized. A similar setting happens also with organization names. A possible workaround would be to use a PoS tagger to aid candidate NE extraction during the generation of the training corpus. Another possibility is to consider a word n-gram based extraction strategy that does not fully focus on capitalized words [13]. In order to reduce the computational effort, the n-gram based strategy could focus only in the extracted sequences of capitalized words, because NEs of the three categories in question should be always capitalized in Albanian. However, we noticed that some of the false positives were caused by a wrongful word capitalization (not following the Albanian spelling rules) in the news articles in question. A possible improvement may also be achieved by extending the gazetteer with foreign names extracted from Wikipedia in other languages.

6. Conclusion

In this work we describe our approach on creating the first automatically generated NE annotated corpus for Albanian. Previous attempts of Albanian NER showed that the state-of-the-art NER supervised methods work well for Albanian; however the corpora used for training had a modest size.

There are no publicly available NLP tools for Albanian (i.e., PoS taggers) that would facilitate the corpus generation. Because of this we had to rely on trivial features for candidate NE selection like sequences of capitalized words. We noticed that this resulted in a lower recall for our automatically generated training corpus when evaluated against human annotated testing data. This was true especially for the person NE class.

We also generated the first Albanian NE gazetteer automatically extracted from Wikipedia. WP article titles were identified as named entities (person, location, organization) using the available WP metadata (categories and intrawiki links) and a manually defined Wikidata ontology mapping. It contains 52,910 named entities in total. This gazetteer was used to automatically annotate a corpus of news articles published in Albanian.

The annotated corpus is exported in Apache OpenNLP format. It contains about 25,000 annotated sentences (with at least an annotated entity) for each NE category that was covered (person, location, organization). The used sentence tokenizer did not perform very well for Albanian. Therefore, not all sentences have been detected correctly. To the best of our knowledge, there is no Albanian training

corpus for a machine learning based tokenizer, and there is also no rule based tokenizer available, so this remains for future works.

As stated in related works [5, 7, 9], automatically generated NE annotated corpora usually is considered of silver quality. However, when there is a lack of human annotated ones, it is possible to use silver corpora even though the achieved recall is usually lower. This was also confirmed by the evaluation results of our corpus. Another possible use of such a corpus is to use it for aiding the annotation process of human annotators, reducing so the annotation time and cost.

References

1. Sang, E. F., F. De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. – In: Proc. of 7th Conference on Natural Language Learning at HLT-NAAC, Association for Computational Linguistics, Vol. **4**, 2003, pp. 142-147.
2. Bender, O., F. J. Och, H. Ney. Maximum Entropy Models for Named Entity Recognition. – In: Proc. of 7th Conference on Natural Language Learning at HLT-NAAC, Association for Computational Linguistics, Vol. **4**, 2003, pp. 148-151.
3. Kono, G., K. Hoxha. Named Entity Recognition in Albanian Based on CRFs Approach. – In: Proc. of 2nd International Conference on Recent Trends and Applications in Computer Science and Information Technology, CEUR-WS.org, Vol. **1746**, 2016, pp. 47-52.
4. Rao, D., P. McNamee, M. Dredze. Entity Linking: Finding Extracted Entities in a Knowledge Base. – In: Multi-Source, Multilingual Information Extraction and Summarization, Berlin, Heidelberg, Springer, 2013, pp. 93-115.
5. Arapakis, I., L. A. Leiva, B. B. Cambazoglu. Know Your Onions: Understanding the User Experience with the Knowledge Module in Web Search. – In: Proc. of 24th ACM International on Conference on Information and Knowledge Management, 2015, pp. 1695-1698.
6. Toral, A., R. Munoz. A Proposal to Automatically Build and Maintain Gazetteers for Named Entity Recognition by Using Wikipedia. – In: Proc. of EACL, 2006, pp. 56-61.
7. Nemeskey, D. M., E. Simon. Automatically Generated NE Tagged Corpora for English and Hungarian. – In: Proc. of 4th Named Entity Workshop, 2012, pp. 38-46.
8. Skënduli, M. P., M. Biba. A Named Entity Recognition Approach for Albanian. – In: Proc. of International Conference on Advances in Computing, Communications and Informatics (ICACCI'13), 2013, pp. 1532-1537.
9. Richman, A. E., P. Schone. Mining Wiki Resources for Multilingual Named Entity Recognition. In: – In: Proc. of ACL, 2008, pp. 1-9.
10. Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives. DBpedia: A Nucleus for a Web of Open Data. – In: Proc. of 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, Springer-Verlag, 2007, pp. 722-735.
11. Attardi, G., V. Cozza, D. Sartiano. Adapting Linguistic Tools for the Analysis of Italian Medical Records. – In: Proc. of 1st Italian Conference on Computational Linguistics CLiC-it & the 4th International Workshop EVALITA, 2014, pp. 17-22.
12. Attardi, G., V. Cozza, D. Sartiano. Annotation and Extraction of Relations from Italian Medical Records. – In: Proc. of 6th Italian Information Retrieval Workshop, CEUR-WS.org, Vol. **1404**, 2015.
13. Prokofyev, R., G. Demartini, P. Cudré-Mauroux. Effective Named Entity Recognition for Idiosyncratic Web Collections. – In: Proc. of 23rd International Conference on World Wide Web, 2014, pp. 397-408.
14. McCallum, A., W. Li. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. – In: Proc. of 7th Conference on Natural Language Learning at HLT-NAACL, Association for Computational Linguistics, Vol. **4**, 2003, pp. 188-191.

15. V r a n d e č i ć, D., M. K r ö t z s c h. Wikidata: A Free Collaborative Knowledgebase. – Communications of ACM, Vol. **57**, 2014, pp. 78-85.
16. D a v i s, M., L. I a n c u. Unicode Text Segmentation. – Unicode Standard Annex, Vol. **29**, 2012.
17. R a t c l i f f, J. W., D. E. M e t z e n e r. Pattern Matching: The Gestalt Approach. – Dr Dobbs Journal, Vol. **13**, 1988.
18. K o h a v i, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. – In: Proc. of International Joint Conference on Artificial Intelligence (IJCAI'95), Vol. **14**, 1995, No 2, pp. 1137-1145.

Received 30.06.2017; *Second Version 17.11.2017*; *Accepted 01.12.2017*