

# A Named Entity Recognition Approach for Albanian

Marjana Prifti Skënduli

Department of Computer Science  
University of New York in Tirana, Albania  
University of Greenwich, London, UK  
marjanaprifti@unyt.edu.al

Marenglen Biba

Department of Computer Science  
University of New York in Tirana, Albania  
marenglenbiba@unyt.edu.al

**Abstract**—Named Entity Recognition (NER) deals with identifying personal, geographical, organizational or other entity types in a raw text. In this paper we propose the first NER model for the Albanian language. Our model is based on the maximum entropy approach. We manually annotate a corpus in the historical and political domains and train the models to generate classifiers that are able to recognize relevant entities in the text. We achieve good performance for precision and recall on the selected domains, despite the scarcity of Albanian corpora and the fact that this paper marks the first NER research for the Albanian language. Experiments demonstrate that the models can be further improved if richer training corpus is provided.

**Keywords**—named entity; recognition; machine learning; natural language processing; Albanian.

## I. INTRODUCTION

Named Entity Recognition (NER) has increasingly gained attention in Information Extraction (IE) and Information Retrieval (IR) and Machine Translation. It is a high level task which comprises knowledge about Part-of-Speech (POS) tagging, phrase and word morphology. The basic idea of NER, initially introduced in [1], is to recognize and classify every linguistic pattern (word) in a structured or unstructured text, into some predefined categories like person name, location name, organization name (ENAMEX), date and time expressions (TIMEX), number, monetary amounts and percentages (NUMEX). Detecting such Named Entities (NEs) becomes a continuously evolving challenge because of language evolution and enrichment, the existence of a wide diversity among language families, diversity of domains, textual genres and entity types.

Considerable research has been dedicated to NER for the English language. Since the 90's these previous works provide a large literary heritage which inspires and guides related research for other languages. German language has been as well studied [3] as well as Spanish and [2]. Approaches for Chinese have been presented in [4, 5, and 6] for Greek in [7], for Italian [8, 9] and many other languages [10].

Up until now, the NER task has been approached using rule-based, knowledge-based and machine learning methods. The majority of work in the NER area proves that there are two principal ways of building NER systems: linguistic grammar-based and statistical models. Hand-crafted grammar-based systems typically obtain better precision, but they require advanced computational linguistic skills. On the other hand, statistical NER systems require building large

collections of texts with linguistic annotations known as corpus, which represents manually annotated training data [11]. NER systems exploit two ways of achieving linguistic pattern evidence: internal evidence deriving from the word and /or word string itself, and external evidence deriving from thorough context.

Almost all NER systems are inclined to suffer the word sense ambiguity problem, which limits them to attain a near-human performance. However, state-of-the-art NER systems for English produce near-human performance. For example, the best system in [12] scored 93.39% of F-measure while human annotators scored 97.60% and 96.95%. These algorithms had roughly twice the error rate (6.61%) of human annotators (2.40% and 3.05%).

In this big picture, to the best of the authors' knowledge there is no evidence for NER approaches or research for the Albanian language, therefore the approach presented in this paper is of interest because Albanian is considered one of the most particular and difficult Indo-European languages. Also it will serve as the basis for future work in this field, yet to be fully explored.

The Albanian language belongs to the extensive Indo-European family. However, thorough linguistics studies show that it has no particular historical affinity to any other language into the Indo-European family, thus it forms a branch of its own: the *Albanian branch*. Albanian is considered a relatively difficult language with complex morphological structure, hence a NER system for Albanian should clearly address the problem of correct identification of NEs (i.e. resolve ambiguities).

In this paper we propose and develop an efficient and robust NER model for Albanian. Several approaches will be analyzed and recent advances will be considered. We show through extensive experiments that our approach is effective and achieves high accuracy even though the dataset has a modest size.

The paper is organized as follows, Section 2 introduces NER and related approaches, Section 3 discusses some particularities of the Albanian language, Section 4 presents the model, Section 5 presents the experiments and we conclude in Section 6.

## II. NAMED ENTITY RECOGNITION

The term *Named Entity Recognition* (NER), which was officially introduced in [11], represents one of the most intensively studied *Information Extraction* (IE) tasks.

Information extraction is defined as the task of finding structured information from unstructured or semi-structured text [13]. It is of great importance to the text mining field and it has been extensively studied in several research communities.

In IE, the main focus is on the event extraction. By definition, the event involves a few named entities, that comprise different semantic classes (e.g. persons, organizations, locations, dates), and some relationships that hold among these named entities. As a result, the two important tasks of IE are named entity recognition and event extraction. However, named entity recognition is fundamental due to the fact that its accuracy predetermines the success rate of the extraction of more complex structures such as events and relations.

#### A. Approaches to NER

Up until now NER task has been approached using rule-based, machine learning and hybrid methods. A considerable number of studies that address NER system development requirements prove that, there are two principal ways of building NER systems: linguistic rule-based and statistical models. Hand-crafted grammar-based systems typically obtain better precision, but they require advanced human expertise in computational linguistics field. On the other hand, statistical NER systems require building large collections of texts with linguistic annotations often referred to as corpus. Recently developed NER systems rely on statistical machine learning methods mostly, but there is a lot to further explore when it comes to maximum entropy models, maximum entropy Markov models, support vector machines and conditional random fields, successfully applied to named entity recognition task.

##### 1) Rule-Based Approaches

The idea behind the rule-based approach consists in being able to extract names using predefined hand-crafted or automatically learned set of rules. Each rule is composed from a pattern and an action. Usually the pattern is a regular expression, whose definition depends on the featured sequence of tokens. On the other hand, the action marks as an entity a sequence of tokens (tokens are produced after the segmentation of the text into linguistic units such as word-like units, punctuation marks, numbers, alphanumerics, utterance boundaries etc.), labelling the start and the end of an entity.

The tokens are characterized from several different features including: the part of the speech tag, the orthography of token (e.g. capitalization), syntax (e.g. word precedence) and dictionaries. Potentially, a sequence of tokens can match with several rules, thus some related policies are defined in order to handle the occurred conflicts. One approach is to order the rules in advance so that they are sequentially checked and fired [13].

NER approaches based on hand crafted rules involve elevated human expertise on linguistic and are labor intensive. The goal of automatic rule learning is achieved through different proposed methods, which are classified in two main groups: top-down and bottom-up approach. In both cases, the most important step is building the training corpus and

manually labelling NEs. Typically in the top-down approach, the next step is about defining general rules that apply to as many as possible training instances. It is a fact that such rules lack precision. However, the system iteratively reacts and defines more specific rules afterwards. Unlikely, in the bottom-up approach, the existing rule set is reviewed in order to define the new specific rules applicable to the training instances. Only then, specific rules are generalized [13].

In this family of approaches, Grishman introduced in 1995, the NYU systems based on handcrafted rules [14]. These systems relied on manually coded rules and manually compiled corpora. Such models resulted efficient when applied on restricted domains, because they were able to detect complex entities much more easily than learning models did.

In conclusion, rule-based NE systems offer limited portability and robustness due to their strong dependency from the originating domain and language. Adaption to new domains or languages is difficult and moreover a costly and time consuming task, yet to be explored and optimized.

##### 2) Statistical Learning Approaches

Currently, the most dominant approach to developing named entity recognition systems is based on statistical Machine Learning methods. The related algorithms used for this purpose, resolve the named entity recognition task by considering it as a sequence labeling issue. Sequence labeling is a pattern recognition task very common to machine learning. Because of its probabilistic nature it has been widely adopted in several natural language processing tasks like part of speech tagging, named entity recognition and chunking. The work in [13] gives a detailed definition of sequence labeling, suggesting the representation of a given sequence of observations as  $x = (x_1, x_2, \dots, x_n)$  and each observation as a feature vector. The assigned label  $y_i$  to a given observation  $x_i$ , can be naturally predicted through a standard classification which is based exclusively on  $x_i$ . While according to sequence labeling, the label  $y_i$  corresponding to observation  $x_i$ , is deduced from both observation  $x_i$  and from other observations and labels closely spaced in regards to position  $i$  in the sequence.

Sequence labeling applied in named entity recognition treats each word as an observation. It goes along with class labels indicating the named entities and their boundaries within the sequence. A commonly applied notation in such scenarios is the BIO notation. It was first introduced in [15]. According to them every entity type **T** is featured from two labels **B-T** and **I-T**, where the first one indicates the beginning of a named entity and the second one indicates a token inside of the named entity. Additionally, there can be a third type of label **O** notating tokens outside the named entity. Figure 1 depicts a random sentence and its NER sequence labeled according to BIO annotation (**PER** stands for persons and **ORG** for organization):

Mark Zuckerberg Founder & CEO of Facebook Inc.  
 B-PER I-PER O O O B-ORG I-ORG

Fig. 1. NER labels according to the BIO notation

Along with the most distinctive conferences like MUC, ACE, and CONLL there were introduced respective annotation guidelines which currently are widely adopted and applied in several NER projects. Figure 2 presents some random sentences labeled according to MUC-6 annotation guidelines:

"Ford Motor Company profit margin"

<ENAMEX TYPE="ORGANIZATION">Ford</ENAMEX> Motor Company profit margin

"President Barack Obama"

President <ENAMEX TYPE="PERSON">Barack Obama</ENAMEX>

"Bridge over Shkumbin River"

Bridge <ENAMEX TYPE="LOCATION">Shkumbin</ENAMEX> river

Fig. 2. NER labels according to MUC-6

NER systems which are based on Machine Learning, aim to convert the identification task into a classification one. Therefore they generally employ a classification statistical model to satisfy this requirement. Such systems use machine learning algorithms in order to identify and classify nouns into predefined instances of Named Entities: persons, locations, times, organizations etc., according to MUC-s definition on the NER task [10]. Depending on the learning method applied, there are three types of machine learning models used for NER: *Supervised*, *Semi-Supervised* and *Unsupervised Machine Learning* model.

The idea behind *Supervised Learning* (SL) consists of a system able to read a large annotated corpus, to memorize lists of entities, and to create disambiguation rules based on discriminative feature spaces [10]. The main challenge of *Supervised Learning* approach is the requirement of a large annotated corpus (comprising training and test data) which is then used to construct a statistical model. Several languages do not own annotated corpuses and further the costly aspect of creating such resources lead to the other two alternative learning methods: *Semi-Supervised Learning* (SSL) and *Unsupervised Learning* (UL).

The term "Semi-Supervised" has been introduced recently. Bootstrapping is the main technique used from SSL which involves some supervision and an initial set of annotated data to begin the learning process. The system tries to find those instances of entities initially learned, in other similar contexts. The learning process is applied repeatedly as often as new examples are found. The repetition of this process leads to the creation of a large processed corpus with entities.

The *Unsupervised Learning* (UL) model is another important machine learning model, whose goal is to build representations from data without any preceding feedback. Basically, the representations rely on lexical resources and patterns and on statistics computed on large unannotated corpora. Unsupervised Learning is recently applied for NER systems, hence most of them are not fully unsupervised. A lot of efforts should be made to develop fully unsupervised

learning models, in order to exploit their dominant advantage compared to rule based approaches: the ability to be easily ported from one domain or language to another.

Currently, SL techniques are the most frequently employed techniques in NER systems. They include *Hidden Markov Models* (HMM-s), *Decision Trees*, *Maximum Entropy Markov Models* (MEMM), *Support Vector Machines* (SVM-s), and *Conditional Random Fields* (CRF-s). HMM-s, MEMM-s and CFR-s will be briefly described in the following sections due to their importance to this dissertation.

## B. Maximum Entropy Modelling

The maximum entropy modelling technique is quite intuitive, especially when applied for linguistic phenomenon modelling. According to [16], maximum entropy modelling is a framework that integrates information from heterogeneous sources for further classification. The data required for a classification problem is described through a number of features and each of these complex features corresponds to a constraint in the model.

The MaxEnt model applied for a named finder tool such as OpenNLP, calculates the probability values of a word belonging to a class (entity type). Thoroughly speaking, given a raw text file, the probability of each class is obtained for each encountered word. The most probable tag corresponding to a word in a text is defined as the tag that has the highest class conditional probability value.

A maximum Entropy approach models a random linguistic process (in our case) making the common assumption that the selected corpus (sample text) represents at its best the phenomenon we want to model and that the distribution satisfies a given set of constraints. This distribution should be as uniform as possible, meaning to have the highest entropy. In the context of a MaxEnt model, the probability of a certain category  $y$  when  $x$  represents a given context (e.g. a phrase or sentence) can be calculated using the following equations [17]:

$$p(y|x) = \frac{1}{Z(x)} \exp \sum_i \lambda_i f_i(x, y) \quad (1)$$

In this equation  $Z(x)$  is a normalizing constant that satisfies the requirement that  $\sum_y p(y|x) = 1$  for  $\forall x$ . Further  $Z(x)$  is noted below:

$$Z(x) = \sum_y \exp \left( \sum_i \lambda_i f_i(x, y) \right) \quad (2)$$

## III. MODEL ANALYSIS AND DESIGN

### A. Albanian Language

Albanian derives from Illyrian language, spoken from ancient Illyrians that were geographically settled in the Western Balkans. Albanian is currently spoken by 7.6 million Albanians living in Albania, Kosovo, Montenegro, Macedonia and other ethnic grounds, in some Western Europe countries and in North America. Despite its late documentation, the

Albanian language is of great interest to linguists and not only, due to its uncommon and archaic traits.

### B. Automatic Language Processing Issues for Albanian

At first glance, speech units selection appears to be a quite simple process: Morphological analysis divides the text into words; the word in the form of written language, will be defined as a string of characters (symbols) involved between two spaces or between a space and a punctuation mark. However, given that the definition of units should be subject to two criteria: the fragmentation should not be difficult or complex and the units should be coherent and meaningful so that they can facilitate refinements in further levels.

In practice automatic speech recognition encounters many difficulties based on the above definition. Common problems to linguistics consist of: apostrophe, hyphen between words, amalgams, flexions, word formation problems (origin, composition, attachment), etc. In the case of apostrophe, the key question is whether to consider it as a punctuation mark or as part of the word itself. For example we have to consider as a single word the names “*Et’hem*”, “*Mit’hat*”, but we cannot do the same with short unified forms such as “*m’i dërgoi*, *t’i dërgoi*” or with the abbreviation “*ç*” (*ç’është /ç’rëndësi ka ky fakt?*) of the word “*çfarë*”.

As mentioned above there are also some words, often referred to as **amalgams** of two existing units. In these cases it is important to identify the two elements, each of them with a specific syntactical role (*megjithëse*, *mirëdita*, *ndonjëherë*, *meqenëse*), in order to eventually deduct writing rules. The author in [18] states that there is a finite number of known amalgams in the Albanian language. According to her, they can be used to create an exhaustive list of composite words and then refer to them to write productive rules. Additionally, there are some Albanian words (including here nouns, verbs, adjectives, pronouns) which undergo structural changes like in the case of **flexions** – declension or conjugation. Algorithmic processing of flexions is a relatively easy task for the automatic Albanian language processing, due to the fact that noun’s declension and verb’s conjugation rules are finite and exhaustive.

Another complex aspect of the morphological analysis is related to word formative cases. So, prefixes and suffixes which are usually considered polysemous, cause irregularities in the word formation mechanism. Therefore it is quite difficult to compile a closed-end systematic list of such words. For example the word “*anti-njerëzor*” contains the prefix “*anti*” which in this case means “*kundër*”, while it is not the case with the following words: “*antikuar*”, “*antilopë*”, “*antik*” etc.

The spontaneous definition of the morphological unit is more difficult than expected. The above mentioned issues have been on the focus of traditional linguistics since some years. Andre Martinet suggests replacing the concept of the “word” with the “morpheme” which is the minimal meaningful unit [19]. These efforts faced several other issues, therefore no full and reliable text mining methods exist.

The difficulty to develop an operative theory of morphological units is closely related to the language nature,

which is characterized from both regularities and linguistic arbitrariness. However, operative systems for automatic speech recognition are designed according to the above mentioned definition of the word in linguistics. The word is still the premier and basic unit based on which the free text is being tokenized in order to facilitate language processing tasks.

### C. List of Entities with Potential Ambiguities in Albanian

Intrinsically, natural languages are characterized from ambiguous traits which occur in various levels of representation. Generally speaking, there are words, usually naming people, places, organizations, etc., that may have different meanings. Such words can potentially be classified as Person, Organization or Location entity. Being able to correctly map word occurrences in free text with actual entities is the main focus of Named Entity disambiguation.

Language ambiguity can be easily overcome from humans due to their ability to analyse the words and their contextual usage. However, it is still a challenge achieving human-like disambiguation performance at machine level and consequently NER and other NLP related tasks become not trivial ones.

In Albanian likewise in English language, capitalization is a very useful feature which helps the identification of NEs in first place, but later the challenge is to correctly classify them. Correct classification is handled using additional information about the text genre, the context in which words occur, titles or designators (like Mr. , Jr. , Dr. , etc.) that typically precede some NEs. For instance, the Albanian word *Vjosa* can indicate a person name, a location (river) name or a company name (e.g. *Vjosa sh.p.k*). Provided that in Albanian upper case strongly indicates proper names (NNP), we are able to resolve *Vjosa*’s grammatical category, but we need abundant contextual information to perform disambiguation (Fig. 3).

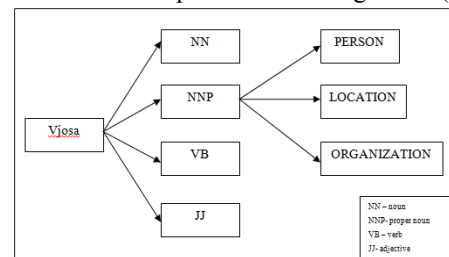


Fig. 3. Example of resolving ambiguity

Common sources of ambiguity in Albanian language are:

1. Geographical units names including rivers, mountains, lakes, cities commonly used as person names: Lura, Vjosa, Drini, Adriatik, Korab, Gramoz, Tomorr, etc.
2. Geographical units names of cities, regions, rivers, mountains, lakes etc. widely used as last names: Shkodra, Elbasani, Korça, Dukagjini, Nivica, etc.
3. Natural phenomenon commonly used as person names: Agimi, Vesa, Drita, Bora, Fjolla, etc.
4. Person names deriving from adjectives: Besnik, Besarta, Bardhoshe, Trim, Mira, Roza, Bardha, Ezmere, Ëmbra, etc.

#### D. Design of the Model

As mentioned previously, Maximum Entropy model is widely known for its ability to combine features from heterogeneous knowledge sources. [20] defines three main components of a ME: futures, histories and features. From there, the calculated probability of history  $h$  to have an outcome  $f$  is equal to  $p(f|h)$ . Following the NER task logic it is important to define the below variables that apply for our Albanian NER model:

$f$  – Named entity class (Person, Location, Organization)  
 $h$  – The information used for weight assignment to features  
 $history$  – the possible context that predicated the class of token  $t$   
 $t$  – token in the corpus

In this scenario  $p(f|h)$  represents the probability of  $f$  associated with token  $t$  in the corpus:

$$p(f|h) = \{f | \text{information relative to token } t (\text{history})\} \quad (3)$$

Features in a ME model are binary values which determine relationships between history and outcome (0 or 1), not simply word attributes. For instance, let  $m$  be a feature and let's assume that the previous POS tag of history  $h$  is a noun. It will mean that token  $t$  has a high probability to be classified as a Person entity.

$$m_j(h_i f_i) = \begin{cases} 1 & \text{if } (POS_{i-1}) \text{ is NN and } t_i = \text{PERSON} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Usually, the *General Iterative Scaling* (GIS) is employed from the model to associate a weight to each feature in the feature space.

### IV. EXPERIMENTAL EVALUATION

#### A. Experimental Setting

We performed our experiments in Apache OpenNLP which is a machine learning based toolkit for the processing of natural language text. It hosts a large variety of java-based libraries which support the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, etc. OpenNLP represents an open source center which contributes with its publicly available tools towards more advanced text processing services [21].

This GIS algorithm implemented in OpenNLP, provides a valuable free resource to the implementation of our NER model for Albanian. The Name Finder tool is able to detect named entities and numbers in a free text, if a model is first trained on that specific entity type and language. For the time being, OpenNLP developers have already made available for demonstration and further testing around 15 pre-trained name finder models trained on existing corpora of Dutch, English and Spanish language only. However, the OpenNLP platform

is flexible enough to allow users build their own corpus in any preferred language and consequently train a model for a specific entity of interest.

#### B. Corpus Building

Prior to building a model, the data must be converted to the OpenNLP name finder training format and encoded as UTF-8. The format requires placing one sentence per line and ensuring that sentences are tokenized by using spans to mark named entities. Below is presented a sample from the training corpus build for this project containing annotated organization entities following the OpenNLP format requirements:

Shqipëria ka marrë angazhim të bashkëpunojë plotësisht me <START.organization> EULEX-in  
 <END> për hetime të plota mbi akuzat e ngritura në rezolutën e miratuar nga  
 <START.organization> Asambleja Parlamentare <END> dhe <START.organization> KiE-së  
 <END> për trajtimin çnjerëzor të njerëzve dhe trafikimin e paligjshëm të organeve në Kosovë,  
 ndërkohë që njëkohësisht i ka hedhur poshtë këto akuza.

Fig. 4. Training corpus sample

In case the corpus will be composed from different documents of the same genre, it is recommended to separate them by empty lines which trigger the reset of the adaptive feature generators [21]. The corpus we built is a collection of three sub-corpus.

The first sub-corpus, namely the “Person Corpus” has been selected so that it can predominantly contain Person entities mainly. The text used to build this sub-corpus is taken from the book “History of Albania” [22]. Its content is made electronically available at <http://historia.shqiperia.com/>. We feel that since this is an academic publication (likely free of orthographic errors), it provides a sound source.

The second sub-corpus, namely the “Location Corpus” has been extracted from the same resource. We took advantage of the fact that some selected chapters of the book “History of Albania” contain extensive descriptions of locations in Albania, including here cities, mountains, rivers etc. In absence of typical geographical genre texts in Albanian language, we thought that nothing better than this resource could help us customize a NER model for Albanian location names. The characteristics of this sub-corpus rate it at the size of 1115 sentences, meaning 33.800 words.

The third sub-corpus should accommodate our need to build a model that can accurately classify organization named entities in Albanian language texts. It was definitely challenging finding an adequate text resource. Thus we decided to change the genre of the text and use some politics related text resources. The sub-corpus was built using some strategic documents, namely the “Stabilisation and Association Reports” publicly available on the Ministry of Integration website <http://www.mie.gov.al/>. Thanks to the nature and context of these political reports, we could build a rich corpus in coherent organization names with 821 sentences and 18.400 words.

#### C. Experimental Results

The experiments were performed using several helpful resources. Among them the most important one is the OpenNLP MaxEnt Java package (version 1.5.2). In this



maximum entropy package, the user should set two parameters: the number of iterations for GIS and the so-called cutoff parameter that shows the minimum number of times that a feature must be seen during training. We measure the accuracy of our models in terms of Precision, Recall and F-measure [23].

In the first experiment, we conducted a 10-fold cross validation by applying the default parameters (iterations = 100 and cutoff = 5). Results are shown in Table I:

TABLE I. RESULTS IN TERMS OF PRECISION, RECALL AND F-MEASURE.

Class	Person Corpus			Location Corpus			Organization Corpus		
Fold	Prec.	Rec.	F-M	Prec.	Rec.	F-M	Prec.	Rec.	F-M
0	0.83	0.52	0.64	0.81	0.65	0.72	0.64	0.50	0.56
1	0.84	0.55	0.66	0.89	0.65	0.75	0.79	0.67	0.73
2	0.90	0.88	0.89	0.90	0.84	0.87	0.71	0.52	0.60
3	0.90	0.82	0.86	0.86	0.75	0.80	0.68	0.58	0.63
4	0.92	0.70	0.80	0.92	0.69	0.79	0.63	0.65	0.64
5	0.95	0.65	0.77	0.53	0.53	0.53	0.79	0.69	0.74
6	0.63	0.71	0.67	0.73	0.56	0.64	0.69	0.42	0.52
7	0.87	0.77	0.82	0.80	0.63	0.71	0.68	0.67	0.68
8	0.82	0.65	0.73	0.86	0.72	0.79	0.71	0.70	0.70
9	0.82	0.75	0.79	0.97	0.62	0.76	0.60	0.55	0.57
Avg.	0.85	0.70	0.76	0.83	0.66	0.73	0.69	0.60	0.64

In the second experiment we aimed to test the model performance for each entity type with a different number of iterations and cut-off values. Results are reported in Table II. The average performance of the model for each entity class increased when the smallest value of cutoff (cutoff=3) was applied.

TABLE II. DIFFERENT NUMBER OF ITERATIONS AND CUT OFF VALUES

Class	Person			Location			Organization		
	F-Measure (i=500)			F-Measure (i=500)			F-Measure (i=100)		
Fold	Cut off 3	Cut off 5	Cut off 7	Cut off 3	Cut off 5	Cut off 7	Cut off 3	Cut off 5	Cut off 7
0	0.96	0.68	0.87	0.75	0.75	0.73	0.61	0.59	0.65
1	0.75	0.76	0.75	0.87	0.85	0.86	0.62	0.63	0.59
2	0.82	0.84	0.84	0.68	0.67	0.66	0.72	0.65	0.65
3	0.74	0.75	0.70	0.85	0.80	0.83	0.65	0.60	0.62
4	0.73	0.69	0.68	0.82	0.77	0.81	0.64	0.63	0.63
Avg.	0.80	0.74	0.77	0.79	0.77	0.78	0.65	0.62	0.63

## V. CONCLUSIONS

In this paper we have presented the first NER model in literature for the Albanian language along with the creation of a tagged corpus of around 3000 sentences and 87900 words. The model is based on a statistical learning approach based on *Maximum Entropy Markov Model*. Through extensive experiments on a corpus composed of texts belonging to the historical-political domain we showed the effectiveness of the model achieving high results considering the modest size of the corpus. As future work we intend to enlarge the corpus in size and in nature including other domains.

## REFERENCES

- [1] Grishman, R. and Sundheim, B. (1996). A brief history. Proceedings of the 16th International Conference on Computational Linguistics. *Message Understanding Conference*, (pages. 466-471).
- [2] CoNLL, (2002). Proceedings of the Conference on Computational Natural Language Learning. Taipei, Taiwan.
- [3] CoNLL, (2003). Proceedings of the Conference on Computational Natural Language Learning. Edmonton, Canada.
- [4] Chen, H. H.; Lee, J. C. (1996). Identification and Classification of Proper Nouns in Chinese Texts. In *Proceedings of the International Conference on Computational Linguistics*.
- [5] Wang, Liang-Jyh; Li, W.-C.; Chang, C.-H. (1992). Recognizing Unregistered Names for Mandarin Word Identification. In *Proceedings of the International Conference on Computational Linguistics*.
- [6] Yu, Shihong; Bai S.; Wu, P. (1998). Description of the Kent Ridge Digital Labs System Used for MUC-7. In *Proceedings of the Message Understanding Conference*.
- [7] Boutsis, S., Demiros, I., Giouli, V., Liakata, M., Papageorgiou, H., Piperidis, S. (2000). A System for Recognition of Named Entities in Greek. In *Proceedings of the International Conference on Natural Language Processing*.
- [8] Black, W. J., Rinaldi, F., Mowatt, D. (1998). Facile: Description of the NE System used for Muc-7. In *Proceedings of the Message Understanding Conference*.
- [9] Cucchiarelli, A., Velardi, P. (2001). Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence. *Computational Linguistics* 27:1.123-131, Cambridge: MIT Press.
- [10] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1): 3-26.
- [11] MUC-6, (1995). *Proceedings of the 6<sup>th</sup> Conference on Message Understanding*, MUC 1995, Columbia, Maryland, USA, November 6-8, 1995, ISBN 1-55860-402-2
- [12] Chinchor, N., & Robinson, P. (1997). MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*.
- [13] Jiang, J. (2012). Mining Text Data. In C. C. Aggarwal, & C. Zhai. Springer.
- [14] Grishman, R. (1995). The NYU system for MUC-6 or Where's the syntax. In *Proceedings of the Sixth message understanding conference MUC-6*. Morgan Kaufmann Publishers.
- [15] Ramshaw, L. and Marcus, M. (1995). Text chunking using Transformation-Based learning. In Yarovsky, D. and Church, K., editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82-94, Somerset, New Jersey. Association for Computational Linguistics.
- [16] Manning, C. D., Schütze H. (1999). "Foundations of statistical natural language processing." Cambridge, MA: MIT Press.
- [17] Berger, A. L., Della Pietra, S. D., and Della Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39-71.
- [18] Lagji, K. (2008). *Hyrje në gjuhësinë kompjuterike*. Tiranë:SHBLU.
- [19] Martinet, A. (1960). *Eléments de linguistique générale*, Colin, Paris.
- [20] Ratnaparkhi, A. (1996). A maximum entropy model for Part-of-Speech tagging. In Brill, E. and Church, K., editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Somerset, New Jersey.
- [21] Drost, I., Ingersoll, G., Margulies, B., Goet, T., Kottmann, J., Baldridge, J., Kosin, J., Morton, T., and Silva, W. (2010). Apache's corenlp.
- [22] Academy of Sciences of Albania, Institute of History (2002). *History of Albania, Vol I*. Tiranë: Toena, ISBN 9992716223
- [23] Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 233-240, NY, USA.