Named Entity Recognition in Albanian Based on CRFs Approach

Gridi Kono
Department of Informatics
Faculty of Natural Sciences
University of Tirana
1001 Tirana, Albania
gridi.kono@gmail.com

Klesti Hoxha
Department of Informatics
Faculty of Natural Sciences
University of Tirana
1001 Tirana, Albania
klesti.hoxha@fshn.edu.al

Abstract

Named Entity Recognition (NER) refers to the process of extracting named entities (people, locations, organizations, sport teams, etc.) from text documents. In this work we describe our NER approach for documents written in Albanian. We explore the use of Conditional Random Fields (CRFs) for this purpose. Adequate annotated training corpora are not yet publicly available for Albanian. We have created our own corpus annotated manually by humans. The domain of this corpus is based on Albanian news documents published in 2015 and 2016. We have tested our trained model with two test sets. Overall precision, recall and F-score are 83.2%, 60.1% and 69.7% respectively.

1 Introduction

Named Entity Recognition (NER) is an important tool in almost all Natural Language Processing (NLP) application areas. NLP systems that include some form of information extraction have gained much attention from both the academic and business intelligence community.

Identifying and classifying words of text into different classes is a process defined as named entity recognition (NER) [ZPZ04]. In simple terms, a named entity is a group of consecutive words found in a sentence, and representing entities of the real world such as people, locations, organizations, dates, etc. For instance in the following sentence: "Matteo Renzi is an Italian politician who has been the Prime Minister of Italy since 22 February 2014 and Secretary of the Democratic Party since 15 December 2013.", "Matteo Renzi", "Italy" and "Democratic Party" can be classified as person, location and organization entities, respectively.

In this work we describe a machine learning approach for recognizing named entities in Albanian text documents. The Albanian language lacks of publicly available annotated training corpora for NER. We have created a custom annotated corpus consisting of news articles written in Albanian published in various online news media. The corpus has been created using a custom built web application software that allowed for n-gram based annotation sessions. Experiments were conducted using Standford CRF based NER toolkit[1]. Results were promising despite the small size of the created corpus.

The rest of this paper is structured as follows.

In Section 2 we will present previous works in NER and related approaches. In Section 3 the Conditional Random Fields approach is described. In Section 4 we will describe our corpus and the methodology used for creating it. In Section 5 we will present experiments and their results. Finally, Section 6 concludes the paper.

2 Related Works

NER approaches have been reported since the early 90s. One of the first works has been described by Rau in [Rau91]. This paper describes the idea of a system that extracts and recognizes company names. It relied on handcrafted rules and heuristics.

Since NER is language dependent, many systems have been presented for different languages. In [DBG+00] is described a NER system that recognizes named entities in texts written in Greek. This

[1] http://nlp.stanford.edu/software/CRF-NER.html

approach followed the MUC-7 NER task definition [CR97] with certain adaptions. Entity classes captured in this paper are people, organizations, location names, date and time expressions, and percent and money expressions. This system is based on finite state machine techniques. The achieved precision and recall were 0.86 and 0.81 respectively.

An interesting study by Pathak et al. [PGJ+13] focuses in clinical named entities. It recognizes three types of named entities like Problem, Test and Treatment. In this study, authors proposed an approach which uses domain specific knowledge in the form of clinical features along with textual and linguistic features. The used textual features are stemming, prefix, suffix and orthographical features. The used linguistics features are part-of-speech(POS), chunks and NP Head. While the used clinical features are section headers, customized stop words, dictionary search, abbreviations and acronyms. They performed experiments with i2b2 shared task using CRF++[2]. The evaluation task was done using micro-averaged precision, recall, and F-Score for exact and inexact matches. For exact matches they achieved 0.889 precision, 0.813 recall and 0.849 F-score respectively. For inexact matches they achieved 0.966 precision, 0.883 recall and 0.923 F-Score.

An approach for German language is presented by Faruqui et al. in [FPS10]. Their work consists of training an existing Stanford NER system on various German semantic generalization corpora. Semantic generalization refers to acquiring semantic similarities from large, unlabelled corpora that can support the generalization of predictions to new, unseen words in the test set while avoiding over-fitting. Corpora was evaluated on both in-domain and out-of-domain data, assessing the impact of generalization corpus size and quality. The F-score of this system improves by 6% (in-domain) and 9% (out-of-domain) over supervised training approaches.

Benajiba et al. in [BDR+08] have developed a NER system for Arabic language. The features used are contextual, lexical, morphological, geographical dictionaries (gazzeters), Part-of-speech tags and Base-phrase-chunking, nationality and the corresponding English capitalization. The system has been evaluated using ACE Corpora[3] and ANERcorp[4]. The aggregate F-score for this system (when all the features are considered) is 82.71%.

A valuable approach for Albanian Language is presented for the first time by Skënduli and Biba in [SB13]. Their work uses a human annotated corpus. The domain of this corpus is focused in Politics and History documents. The corpus is a collection of three sub-corpora: People corpus, Locations corpus and Organizations corpus. They performed experiments with these corpora using Apache OpenNLP[5] as a framework for running their machine learning based NER approach. The achieved results of this approach were as follows:

The People corpus produced values of Precision, Recall and F-score as 0.85, 0.70 and 0.76 respectively. The Locations corpus produced values of Precision, Recall and F-score as 0.83, 0.66 and 0.73 respectively. While Organizations corpus produced values of Precision, Recall and F-score as 0.69, 0.60 and 0.64 respectively.

In general, NER approaches reported for most languages belong to these categories:

1. Rule Based
2. Machine Learning
3. Hybrid Models

The first one is based on handcrafted rules, linguistic approaches and Gazzeters. The second is based on statistical methods. The most used methods for statistical NER are Maximum Entropy Model [SB13], Conditional Random Fields [PGJ+13,FPS10,LMP01], Hidden Markov Models [ZS02] and Support Vector Machines [BDR+08]. The third one combines Rule based and Machine learning methods [Rau91, DBG+00, BDR+08]. Machine learning based methods depend on preliminary training. The training methods can be divided into three groups: Supervised learning, Semi-supervised learning and Unsupervised learning method. Supervised methods need annotated training data to retrieve optimal results from the classifier. Semi-supervised learning methods require some data which are used as a help for the training. Unsupervised learning methods do not depend on training data and are mostly clustering based.

3 Conditional Random Fields

In this work we used a linear chain CRF sequence classifier. Conditional Random Fields is a probabilistic framework used to segment and label sequence data. Conditional Random Fields are undirected graphical models, used to calculate the conditional probability of values on designated output nodes, given already assigned values to the input nodes. The conditional probability of a state sequence $\mathbf{x} = (x_1, \ldots, x_T)$ given an observation sequence $\mathbf{y} = (y_1, \ldots, y_T)$ calculated as:

[2]https://taku910.github.io/crfpp/
[3]http://corpus.ied.edu.hk/ace/Corpus.html
[4]http://users.dsic.upv.es/ ybenajiba/
[5]https://opennlp.apache.org/

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}\theta_k f_k(y_{t-1}, y_t, x_t)\right\} \quad (1)$$

where, $f_k(y_{t-1}, y_t, x_t)$ is a feature function whose weight θ_k , is to be learned via training. The values of feature functions may range between $-\infty$ to $+\infty$, but usually they are binary. Usually, when applying CRFs to the named entity recognition problem, an observation sequence is a sequence of tokens or a raw text and the state sequence is its corresponding sequence of labels [LMP01]. By Hammersley-Clifford theorem, the conditional probability of a state sequence given an input sequence will be:

$$Z_\theta(\mathbf{x}) = \sum_{\mathbf{y}\in Y^T} \exp\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}\theta_k f_k(y_{t-1}, y_t, x_t)\right\}. \quad (2)$$

where Z_θ is a normalisation factor over the all state sequences, which ensures that the probability distribution sums up to 1.

4 Corpus

There are no publicly available NER annotated corpora for Albanian texts. Hence we decided to create a corpus of Albanian based on news articles published online from different local newspapers. We have used the news aggregator for Albanian news, built by [HBN16] using Scrapy[6]. News articles retrieved by this news aggregator are stored in a MySQL database. We used Python NLTK toolkit to generate all n-grams(for n=1,2,3,4) for each news article. All generated n-grams are stored in the same database with corresponding news articles. In this paper we have considered only unigrams. Figure 1 shows the workflow diagram of building our corpus.

[6]https://scrapy.org/

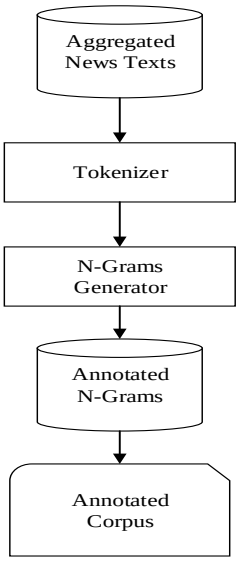

Figure 1: Workflow diagram.

In order to add labels to the generated unigrams, we have built a web application using ASP.NET WebForms[7], C#[8], JQuery/Ajax[9] and Javascript technologies.

Our application has two simple user interfaces. The first user interface (Figure 2) lists titles of news articles and allows selecting each of them for n-gram labeling.

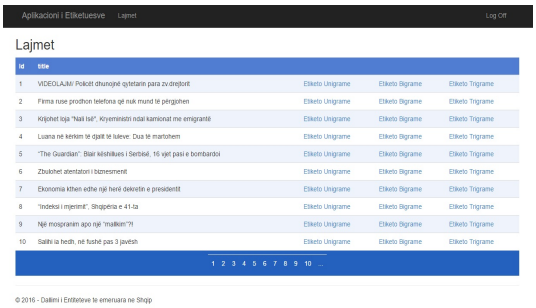

Figure 2: News User Interface.

The second user interface (Figure 3) consists of two parts. The first part displays raw content of a selected news article and the second part displays all unigrams of it. For each unigram, annotators are able to set a corresponding label from a list of predefined entity classes. Actually, our web application offers interfaces for also labeling bigrams and trigrams, but because the NER training model that we used for our experiments depends on labeled unigrams we were limited to these.

In order to visually aid the entity identification process, each word which starts with an uppercase charac-

[7]https://www.asp.net/web-forms/
[8]https://msdn.microsoft.com/en-us/library/67ef8sbd.aspx
[9]https://jquery.com/

Figure 3: Unigrams User Interface.

ter inside the news content is highlighted with yellow color.

This web application allows annotators to work on the same news item without overriding previous n-gram labels, but storing each annotation instead allowing so for quality control processes. However, we avoided this for the experiments reported in this work, leaving the experimentation with annotation quality assurance techniques for future works.

Our corpus consists of 130 documents. The selected news documents were published in two different years (2015 and 2016). They belong to eight categories: Politics News, Economic News, Sport News, Health News, Technology News, Culture News, Chronics and Opinions.

This corpus has been manually annotated by humans. We have organized three sessions with volunteer annotators in order to annotate more n-grams. In the first and second sessions, volunteers annotated all news articles designated for the training set. In the third session we used different annotators that have not participated in previous sessions, in order to annotate test sets. The annotation has been done according to the Inside Outside(IO) format[10] with four tags as described in Table 1.

NE tag	Meaning	Example
PER	person name	George PER Bush PER
LOC	location name	Tirana LOC
ORG	organization name	OSCE ORG
O	Not an entity	76% O

Table 1: Named Entity Tagset

5 Experiments and Results

5.1 Experimental Set-up

We performed our experiments in Stanford NER. Stanford NER is a Java implementation of a Named Entity Recognizer. It comes with well-engineered feature extractors for Named Entity Recognition, and many options for defining feature extractors. Stanford NER is also known as CRFClassifier. The software provides a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models. That is, by training your own models on labeled data, you can actually use this code to build sequence models for NER or any other task [FGM05].

[10] http://nlp.stanford.edu/software/crf-faq.html

5.1.1 Evaluation Metrics

We have evaluated the results of our experiments with well-accepted standard measures in evaluation of trained NER models. This can be performed by annotating a corpus and then compare the human annotations with a gold standard corpus. Thus, each annotation must be classified as being a:

1. True Positive (TP): the system provides an annotation that exists in the gold standard corpus.

2. True Negative (TN): the non existence of an annotation is correct according to the gold standard corpus.

3. False Positive (FP): the system provides an annotation that does not exist in the gold standard corpus;

4. False Negative (FN): the system does not provide an annotation that is present in the gold standard corpus.

Concretely we used Precision, Recall and F-score as used by other authors in [DBG+00] [BDR+08] [SB13].

Recall measures the ability of a NE trained model to present all relevant entities, and is formulated as:

$$Recall = \frac{TP}{TP + FP}$$

Precision measures the ability of a NE trained model to present only relevant entities, and it is formulated as:

$$Precision = \frac{TP}{TP + FN}$$

These two measures of performance can be combined as one performance metrics, the F-score, which is computed by the weighted harmonic mean of precision and recall.

$$F - score = 2 \times \frac{Precision \times Recall}{Precison + Recall}$$

5.1.2 Experiments

Our corpus is further divided into the training and the test set, which contain 100 and 30 documents respectively.

The training set contains news documents published in 2015, in total around 50.000 words.

The test set is divided into two subsets. The first subset contains news documents published in 2015, while the second subset contains news documents published in 2016. Each subset contains 15 documents respectively.

We have conducted two experiments, the first using the first subset of test data and the second makes use of the second subset.

6 Results

The evaluation task for each experiment as described above was done using three different metrics: Precision, Recall and F-score. The following tables show results for each test set that has been used. The used training model is the same for both experiments. These calculations were carried out automatically by Stanford NER.

In the first experiment the NE class with highest F-score is Locations class, 81.1%. The NE class with lowest value is Organizations class, 47.1%. Overall for the first experiment we have got Precision of 80.8%, Recall of 64.0% and F-score of 71.4% (see Table 2).

Entity class	Precision	Recall	F-score
Locations	0.8219	0.8000	**0.8108**
Organizations	0.6154	0.3810	**0.4706**
People	0.8409	0.5441	**0.6607**
Average	0.8077	0.6402	**0.7143**

Table 2: Results for the first experiment.

In the second experiment the NE class with highest F-score is People class, 78.7%. The NE class with lowest value is Organizations class, 35.3%. Overall for the first experiment we have got Precision of 85.6%, Recall of 56.3% and F-score of 67.9% (see Table 3).

Entity class	Precision	Recall	F-score
Locations	0.8706	0.6379	**0.7363**
Organizations	0.8333	0.2239	**0.3529**
People	0.8429	0.7375	**0.7867**
Average	0.8555	0.5627	**0.6789**

Table 3: Results for the second experiment.

The overall average Precision, Recall and F-score are 83.2%, 60.1% and 69.7% respectively (see Table 4).

	Precision	Recall	F-score
Experiment I	0.8077	0.6402	0.7143
Experiment II	0.8555	0.5627	0.6789
Average	**0.8316**	**0.60145**	**0.6966**

Table 4: Final results of experiments.

7 Conclusions and Future Directions

In this paper we presented the results of a machine learning approach for identifying named entities in text documents written in Albanian. It is based in Conditional Random Fields and was evaluated against two different test sets on a corpus of Albanian news documents. The corpus was created by annotating news articles through the use of a custom built web application software. Volunteer annotators manually performed this process by using a n-gram based news visualization interface. The experiments were restricted in the recognition of three entity classes: people, locations, and organizations.

Even though the size of the annotated corpus is modest, we got promising results, showing that the experimented model can be used for successfully extracting named entities from Albanian text documents. The relatively low recall values for organization entities may be improved by using a larger corpus and expand it beyond news text documents written in Albanian.

In the future we intend to increase the size of the corpus in order to get more significant results. Furthermore, we aim to improve the quality of the annotated data by switching to a semiautomatic corpus creation approach [ACS14]. It would need to use a publicly available knowledge base of people, locations, and organizations. This way we may aid human annotators in better recognizing possible named entities in the provided texts. Also we want to improve the user interface involved in the annotation process and also tweak it in order to avoid confusion and produce annotation results better suited for the NLP toolkit that is being used. Another aspect that we want to improve in the future, is the inclusion of a quality control scheme in the annotation process. This way we will be able to avoid false or ambiguous tagging of named entities present in the text documents in question.

Experimenting with other NER machine learning techniques like Hidden Markov Model (HMM), Support Vector Machine (SVM) and studying the behaviour of these approaches for Albanian written documents is also in our future plans.

A NER tool for Albanian texts will also enable concrete applications like the creation of a knowledge base that stores facts about named entities present in news articles [HBN16].

References

[ACS14] Giuseppe Attardi, Vittoria Cozza, and Daniele Sartiano. Adapting linguistic tools for the analysis of italian medical records. In *Proceedings of the First Italian Conference on Computational Linguistics*, 2014.

[BDR+08] Yassine Benajiba, Mona Diab, Paolo Rosso, et al. Arabic named entity recognition: An svm-based approach. In *Proceedings of 2008 Arab International Conference on Information Technology (ACIT)*, pages 16–18, 2008.

[CR97] Nancy Chinchor and Patricia Robinson. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, page 29, 1997.

[DBG+00] Iason Demiros, Sotiris Boutsis, Voula Giouli, Maria Liakata, Harris Papageorgiou, and Stelios Piperidis. Named entity recognition in greek texts. In *LREC*, 2000.

[FGM05] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[FPS10] Manaal Faruqui, Sebastian Padó, and Maschinelle Sprachverarbeitung. Training and evaluating a german named entity recognizer with semantic generalization. In *KONVENS*, pages 129–133, 2010.

[HBN16] Klesti Hoxha, Artur Baxhaku, and Ilia Ninka. Bootstrapping an online news knowledge base. In *International Conference on Web Engineering*, pages 501–506. Springer, 2016.

[LMP01] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.

[PGJ+13] Parth Pathak, Raxit Goswami, Gautam Joshi, Pinal Patel, and Amrish Patel. Crf-based clinical named entity recognition using clinical nlp. In *Proceedings of 10th International Conference on Natural Language Processing*, 2013.

[Rau91] L. F. Rau. Extracting company names from text. In *Proc. Seventh IEEE Conf Artificial Intelligence Applications*, volume i, pages 29–32, February 1991.

[SB13] Marjana Prifti Skënduli and Marenglen Biba. A named entity recognition approach for albanian. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pages 1532–1537. IEEE, 2013.

[ZPZ04] Li Zhang, Yue Pan, and Tong Zhang. Focused named entity recognition using machine learning. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 281–288, New York, NY, USA, 2004. ACM.

[ZS02] GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480. Association for Computational Linguistics, 2002.