A brief Survey on Named Entity Recognition in Amazighe language

Meryem Talha, Siham Boulaknadel, Ahmed Hammouch

Abstract—Named Entity Recognition (NER) is a core subtask of Information Extraction (IE), the main idea is to extract named entities from a given text and classify them into a set of named entity classes such as person names, locations and organizations, etc. It is an essential part in many applications, such as Machine Translation (MT) and Question Answering System (QA). Amazighe NER has gained considerable attention in the past few years, however, the peculiarities of Amazighe arise the challenges of this task. In this paper, a survey regarding the recent progress made in Amazighe NER research using different techniques is presented.

Index Terms— Named Entity Recognition, Amazighe Language, Machine learning, SVM, rule-based approach, hybrid approach, Corpus.

1 Introduction

Identifying and classifying entities of a given text into different predefined classes, such as names of persons, organizations, locations, expression of times, quantities, monetary values, temporal expressions, percentages, etc., is a process called named entity recognition (NER) [1]. The term Named Entity was first introduced in the sixth Message Understanding Conference (MUC-6), this term covers not only proper names but also includes temporal expressions, numerical expressions and other types of units depending on domain of interest.

In MUC-6, Named entities (NEs) were classified into three types of category as follows [2]:

- ENAMEX: person, organization, location
- TIMEX: date, time
- NUMEX: money, percentage, quantity

NER also provides basic inputs for various NLP tasks, including:

- Information Extraction,
- Questions Answering,
- Automatic Summarization,
- Machine Translation,

A much of NER work has been done in English and some other foreign languages similar to Spanish, French, Chinese, Arabic, etc., with great precision but NER in Amazighe language is at a primary stage.

This article investigates the progress in Amazighe NER research. To our best knowledge, Amazighe NER and categorization have not yet been surveyed, which has motivated us to conduct this survey.

The remaining sections of this survey are organized as follows: Section 2 provides background information about NER approaches, section 3 describes the specificities of the Moroccan Amazighe language, section 4 discusses the linguistic issues and challenges impeding the extraction of named entities in Amazighe language, section 5 reports the Amazighe linguistic resources that have been used for the Amazighe NER task, a state-of-the-art in Amazighe NER research is presented is section 6 and the survey is concluded in section 7.

2 NER approaches

NER systems have been developed using mainly three approaches: the rule-based approach, the machine-learning based approach and the hybrid approach.

2.1 Rule-Based approach

The rule-based approaches typically make use of two types of resources [3]:

- Handcrafted rules set manually written by linguists.
- A set of gazetteers including the lexical trigger words.

The main advantage of the rule-based NER approaches is that they are able to detect complex entities due to the solid linguistic knowledge; however the main disadvantage is the non-ability of portability, these approaches give better results on restricted domains and in order to apply them on new ones, it requires new adapted grammatical knowledge and background of the particular domain. However, for low-resourced languages, handcrafted rules remain the preferred technique.

2.2 Machine-Learning approach

Much of the current researches in NER on the well-known language involve machine-learning based approaches. These approaches treat NER as a classification process and make use of a large amount of NE annotated training data [4].

There are two types of machine learning models that are used for NER called Supervised (SL) and Unsupervised (UL) machine learning model. The main difference between these types is that the first one requires the availability of large an-

- *Meryem Talha is currently a PhD student at the Faculty of Science, Mohammed V University in Rabat, LRIT-CNRST (URAC No. 29), Morocco. E-mail: meriem.talha@gmail.com*
- *Siham Boulaknadel is currently researcher at the Royal Institute of Amazighe Culture Allal El Fassi Avenue, Madinat Al Irfane, Rabat-Instituts, Morocco. E-mail: boulaknadel@ircam.ma*
- *Ahmed Hammouch is currently the Head of the Department of Scientific and Technical Affairs in the National Centre for Scientific and Technical Research (CNRST), Morocco. E-mail: a.hammouch@cnrst.ma*

notated data in the training stage while the second one does not need an annotated data beforehand; it relies on clustering similar documents or entities together.

Several ML techniques have been widely used for the NER task of which hidden markov model [5], maximum entropy [6], conditional random field [7], Support Vector Machine [8] are most common.

Unlike the rule-based approaches, Machine-learning based approaches can be easily applied to different domain or languages. However, creating large enough training sets for them remains a problem.

2.3 HYBRID APPROACH

The hybrid approach is the combination of rule-based and machine learning-based approaches, the main idea is to make use of the strongest points from each approach and optimize the overall performance [2].

3 Amazighe Language

Amazighe Language belongs to the branch of the large Afro-Asiatic (Hamito-Semitic) linguistic family [9-10]. It covers a boundless geographical zone: all of North Africa, the Sahara (Tuareg), and a part of the Egyptian oasis of Siwa.

In Morocco, Amazighe language is one of the national and official languages besides the classical Arabic. According to the last census of 2014[1], it is spoken by close to 27% of the population. Moreover, three different dialect clusters have been differentiated: Tarifite (4.1%) in North (Rif), Tamazight (7.6%) in Central Morocco (the Mid-Atlas and a part of the High-Atlas) and South-East, and Tachelhite (15%) in the South-West and the High Atlas.

4 Linguistic Issues and Challenges of Amazighe Language

Amazighe is a highly agglutinative language and it makes Amazighe language morphologically rich with very productive inflectional and derivational processes, and it differs from English or other Indo European languages. Despite the achievements made in Amazighe NER research, the task still remains challenging due to many issues.

For example, this language does not have capitalization, which is a major feature used by NER systems for European languages. Another issue is that the official alphabet for Amazighe language in Morocco is "Tifinagh", which is different from Latin Alphabets (a □; □ b; □ c; □ d). However different scripts have been frequently used to write the Amazighe language such as Latin and Arabic scripts.

Another issue is the lack of available Semantic and Linguistic Resources for Amazighe NER, corpora and lexical resources are the two main types of linguistic resources. Among other problems, one last example is the lack of standardization and spelling variation, the Amazighe text does not respect the standard writing convention. For example, the person name ("□□□□□□□, bnkiran, Benkiran") can be written as ("□□ □□□□□, Bn Kiran, Ben Kiran"), and the Location name ("□□□□ □□ □□□□□, Fkih Ben Saleh") that can be-written as ("□□□□ □□□□□□□, fkihbnsaleh").

5 Amazighe Linguistic Resources

Amazigh is a low-resource language, however with the growing interest in Amazighe NER research, some efforts has been made in creating standardized linguistic resources in order to facilitate the development of Amazighe NER systems. This section discusses the various resources created.

5.1 CORPORA

As mentioned before, the main problem of Amazighe language is the lack of publicly available annotated corpora. Therefore, some researches have built their own corpora for training and testing purposes.

The last updated version of the Amazighe corpus "AMCorp" contains more than 900 news articles published online[2], that are collected from a broad range of topics (sports, economics, news on royal activities of His Majesty King Mohammed VI, and many others), containing news that happened over a period of 2 years (dated between May 2013 and July 2015). The articles are selected in such a way that the data set contains different types of information, and that the system's future use will not limited to any particular text type. It consists of nearly 170.000 words, after some data cleaning operations like deleting non-Amazighe words. This data set is manually annotated following the MUC guidelines, ENAMEX (Location, Person, and Organization), NUMEX (Numbers, Percentage and Money) and TIMEX (dates & times) types.

5.2 GAZETTEERS

As far as the corpus is concerned, the gazetteers are also needed in Amazighe NER systems. They are considered one of the important Amazighe linguistic resources, a description of the Amazighe gazetteers is given below:

- Person gazetteer: consists of famous person names splitted into first name and last name gazetteers in order to identify different combinations. It contains 2533 entries.
- Location gazetteer: includes different location names in Morocco, with names of almost all the countries in the world, cities, states, and geo-graphical names from different sources. The total of entries is 2318.
- Organization gazetteer: contains names of important organizations such as those of political parties, universities, agencies and banks. 913 entries have been collected.
- Date/Time gazetteer: includes entities related the temporal expressions, such as days and months entities. The majority of these entities were extracted from the IRCAM3 website. This contains 193 entries.
- Additional gazetteers of numerical expressions including numbers (216 entries), money (14 entries) and percentage (3 entries) have been created.

[1] http://www.hcp.ma/Presentation-des-premiers-resultats-du-RGPH-2014_a1605.html

[2] http://www.mapamazighe.ma

[3] www.ircam.ma

- Lists of 468 trigger words have also been created manually, which generally provide cues surrounding the named entities that would indicate their presence, it contains titles like (□□□□, Mass, Mr) and (□□□ □ □□□□□□□ □□□□□□□□ □□□□□□ □□□□□, bab n tattuyt tagldant agldun mulay, His Majesty the King).

6 Amazighe NER Systems

In this section we present different Amazighe NER systems. They are classified according to the approaches used. But before, we should mention that all experiments have been done using GATE.

6.1 GATE

A good number of tools are available for developing and evaluating NER systems. We have chosen GATE[4] because it's one of the most popular freely available software tools dealing with NLP. The tool supports nine languages (English, French, German, Italian, Chinese, Arabic, Romanian, Hindi, and Cebuano) [11-12]. It provides a framework which the development of the rule-based NER systems is easy; the user has the ability of implementing grammatical rules as a finite state transducer using JAPE, even the machine-learning based systems can be implemented using GATE.

6.1 RULE-BASED SYSTEMS

One of the first research papers in the field was presented by Talha and al. [13]. The paper describes an Amazighe rule-based NER system. It is able to extract and recognize person names, locations, organizations. It relies on a set of 17 grammar rules and 3710 lexical resources. For evaluation, 200 texts from AmCorp were selected randomly and manually tagged. The overall performance obtained for the various categories: person, location, and organization, was an F-measure of 64%, 40% and 82%.

As a continuation of the initial attempt, Boulaknadel and al. [14] developed an enhanced rule-based Amazighe NER system. The system identifies the following NE types: person names, locations, organizations, date and numbers. For the experiment, the authors used around 289 news, the size of gazetteers was 4666 entries. They reported an f-measure of 83% for person names, 97% for locations, 76% for organizations, 67% for dates and 95% for numbers.

Another NER system adopting the rule-based approach for recognition and classification of Amazighe named entities is presented by Talha and al. [15]. In this research, the authors added some gazetteers (5193) and grammar rules (76) to the system to increase the performance. They applied the set of rules and gazetteers on a corpus containing 430 news. The system was able to recognize five classes of named entities. It obtained an F-measure of 81.5% for person, 87.75% for location, 84% for organization, 80% for date, and 83.5% for numbers.

6.2 MACHINE LEARNING SYSTEMS

The work of Talha and al. [16] is a new attempt to improve performances of the previous Amazighe NER systems.

The authors tried to recognize the Amazighe named entities using a supervised machine learning approach (using SVM [17]) and exploring different sets of features. The features include token form, token kind, semantic classes from gazetteer lists and named entity type.

The system is able to identify the following NE types: person, location, organization, and numbers, Date/Time, Money and Percentage. The overall system's performance in terms of F-measure was as follows: 81%, 82%, 86%, 88%, 94%, 94 and 100%, respectively using training set of 800 texts and test set of 100 texts.

6.3 HYBRID SYSTEMS

Recently, Talha and al. [18] proposed a hybrid NER system for Amazighe. The rule-based component is a duplication of the NERAM system [15]. The ML-based component uses the SVM classifier. The feature set used includes the NE tags predicted by the rule-based component, contextual features and the gazetteers features.

The system identifies the following types of NEs: person names, locations, organizations, dates and numerical expressions. The overall performance obtained for the various categories using AmCorp was an f-measure of 73%. The experimental results showed that the hybrid Amazighe NER approach didn't attempt a very good improvement of results compared to the rule-based and the ML-based components when they are processed individually, this is due to the minimized feature set used, the lack of POS tagging and morphological features.

7 Conclusion

This paper has presented a brief literature review of the major works done regarding the concept of named entity recognition for Amazighe language. Amazighe NER works are in progress, the number of current Amazighe researches is still insufficient compared with the others well-resourced languages due to many issues such as the lack of a huge annotated corpora, lack of capitalization, variations in writing style and difficult morphology. Our main aim is to provide a key to deal in some detail with Amazighe NER research and guides researchers in interesting and fruitful research directions.

References

[1] Nadeau, D.; "Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision," Ottawa Carleton Institute for Computer Science, School of Information Technology and Engineering, University of Ottawa, 2007.

[2] Grishman, Ralph; Sundheim, B. 1996. Message Understanding Conference - 6: A Brief History. In Proc. International Conference on Computational Linguistics.

[3] Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1). pp. 3-26. 2007.

[4] Sharnagat, Rahul. "Named Entity Recognition: A Literature Survey." (2014).

[4] General Architecture for Text Engineering, https://gate.ac.uk/

[5] Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997, March). Nymble: a high-performance learning name-finder. In Proceedings of the fifth conference on Applied natural language processing (pp. 194-201). Association for Computational Linguistics.

[6] Borthwick, A., Sterling, J., Agichtein, E. , and Grishman, R. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. 1998.

[7] McCallum, A., & Li, W. (2003, May). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In the Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 (pp. 188-191). Association for Computational Linguistics.

[8] Y.C. Wu, T.K. Fan, Y.S. Lee, S.J Yen, ``Extracting Named Entities Using Support Vector Machines", Spring-Verlag, Berlin Heidelberg, 2006.

[9] Joseph Greenberg. “The Languages of Africa”. The Hague, 1966

[10] Omar Ouakrim. “Fonética y fonología del Bereber”. Survey at the University of Autònoma de Barcelona, 1995.

[11] Cunningham, Hamish. 2002. Gate, a general architecture for text engineering. Computers and the Humanities, 36(2):223–254.

[12] Maynard, Diana, Hamish Cunningham, Kalina Bontcheva, Roberta Catizone, George Demetriou, Gaizauskas Robert, Oana Hamza, Mark Hepple, Patrick Herring, BrianMitchell, Michael Oakes, Wim Peters, Andrea Setzer, Mark Stevenson, Valentin Tablan, ChristianUrsu, and YorickWilks. 2000. A survey of uses of gate. Technical Report CS-00-06, Department of Computer Science, University of Sheffield.

[13] Talha, M.; Boulaknadel, S.; Aboutajdine, D. NERAM: Named Entity Recognition for Amazighe language. In: 21th International conference of TALN. pp. 517–524. Aix Marseille University, Marseille. 2014.

[14] Boulaknadel, S.; Talha, M.; Aboutajdine, D.; Amazighe Named Entity Recognition Using a Rule Based Approach. In: 11th ACS/IEEE International Conference on Computer Systems and Applications. Doha, Qatar. 2014.

[15] Talha, M. ; Boulaknadel, S. ; Aboutajdine, D. L’apport d’une approche symbolique pour le repérage des entités nommées en langue amazighe. In: EGC. pp. 29–34. Luxembourg. 2015.

[16] Talha, M., Boulaknadel, S., & Aboutajdine, D. (2018, February). Performance Evaluation of SVM-Based Amazighe Named Entity Recognition. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 232-241). Springer, Cham.

[17] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In Machine Learning, pages 273-297, 1995.

[18] Talha, M.; Boulaknadel, S.; Aboutajdine, D. Development of Amazighe Named Entity Recognition System Using Hybrid Method. Journal of Research in Computing Science 90: 151-161. 2015.