Development of Amazighe Named Entity Recognition System Using Hybrid Method

Meryem Talha, Siham Boulaknadel, Driss Aboutajdine

LRIT, Associate Unit to CNRST, Faculty of Science, Mohammed V University Rabat, Morroco
Royal Institut of Amazighe Culture, Allal El Fassi Avenue, Madinat Al Irfane, Rabat-Instituts, Morroco
CNRST, Angle FAR Avenues and Allal El Fassi, Hay Riad, BP 8027 NU, 10102 Rabat, Morocco
meriem.talha@gmail.com, boulaknadel@ircam.ma, aboutaj@fsr.ac.ma

Abstract. The Named Entity Recognition (NER) is very important task revolving around many natural language processing applications. However, most Named Entity Recognition (NER) systems have been developed using either of two approaches: a rule-based or Machine Learning (ML) based approach, with their effectiveness and weaknesses. In this paper, the problem of Amazighe NER is tackled through using the two approaches together to produce a hybrid system with the aim of enhancing in general performance of NER tasks. The proposed system is able of recognizing 5 different types of named entities (NEs): Person, Location, Organization, Date and Number. It was tested on a corpus of Amazigh reports containing 867 diverse articles. Furthermore, a comparison with the baselines of the system based on the case of using just gazetteers and hand-written heuristics is presented. We also provide the detailed analysis of the results.

Keywords: Amazighe Language, Named Entity Recognition (NER), Hybrid Method, GATE

1 Introduction

Named Entity Recognition (NER) is an important subfield of the broader research area in Information Extraction from textual data, aimed at identifying and associating just some types of atomic elements in a given text to a set of predefined categories such as names of persons, organizations, locations, dates, and quantities, called Named Entities (NE)[1]. It serves as the basis for many other crucial areas such as Information Processing & Management[2], financial documents[3], business information documents[4] and biomedical texts[5], particularly involving information retrieval [6]; semantic annotation[7]; classification; ontology population[8]; opinion mining[9], filtering and summarization[10]; question answering[11]; machine translation[12], browsing and visualization; and human-computer interaction in information systems. The term Named Entity

was first used at the 6th Message Understanding Conference (MUC)[13], where the importance of the semantic identification of persons, organizations and localizations, as well as numerical expressions such as time and quantities was obvious. Although the task is given considerable research attention for so many languages including English, French, Spanish, Chinese, and Japanese, etc.

Named entity recognition research on Amazighe texts is known to be scarce. To the best of our knowledge, [14] present the first study on the topic where a rule based named entity recognition system is proposed and evaluated on an Amazighe corpus which contains 200 Amazighe texts, the system was able to extract 3 different types of NEs including Person, Location, Organization. As a continuation of the previous research work, [15] have presented a system which carries out named entity recognition using a set of heuristic rules and lexical resources, they evaluated their system on a corpus containing 289 texts, that can recognize five NE types including Person, Location, Organization, Expressions of Time, Numbers. Lastly, In [16], authors selected 430 Amazighe texts to work on, and they employed a set of lexical resources and sets of rules as information sources, they obtained remarkable results in the detection of Person, Location, Organization, Expressions of Time and Number entities. In this paper, we present a hybrid named entity recognizer for Amazighe texts.

The remainder of the paper is organized as follows. Section 2 presents a background of Amazighe language features illustrating the challenges posed to NER. In section 3 we discuss the details of our approach including system architecture and the machine learning algorithm used, experimental sets and results obtained are shown in section 4. Finally, we discuss the results and some of our insights in section 5.

2 Amazighe Language Features

The Amazighe language, known as Berber or Tamazight, is a branch of the Afro-Asiatic (Hamito-Semitic) languages [17][18]. In Morocco, this language is divided, according to historical, geographical and sociolinguistic factors, into three main regional varieties: Tarifite in North, Tamazight in Central Morocco and South-East, and Tachelhite in the South-West and the High Atlas. However in 2001, thanks to IRCAM[19] efforts, the Amazigh language has become an institutional language nationally recognized; and in July 2011, it has become an official language besides the classical Arabic.

Nowadays, The Tifinaghe-IRCAM graphical system has been adapted in writing Amazighe language for technical, historical and symbolic reasons. It is written from left to right and contains 33 alphabets (27 consonants; 2 semi-consonants and 4 vowels)[20].

2.1 Challenges Undertaken by Amazighe NER

A lot of Named Entity Recognition Systems have been already done thanks to the impulse of MUC conferences. However most of these works have been concentrated on English and other European languages. Yet, named entity recognition

research conducted on Amazighe texts is still rare as compared to related research carried out on other languages.

In particular, Applying NLP tasks to Amazighe are very challenging because of its particularities and unique nature. The main features of Amazighe that pose non-trivial challenges for NER task are as follows:

- No Capitalization: The absence of the uppercase / lowercase distinction represents a major obstacle for the Amazighe language. In fact, the NER for some languages such as Indo-European languages is mainly based on the presence of capital letters which is a very useful indicator to identify proper names in major languages using the Latin alphabet. Uppercase letters, however, do not occur, neither at the beginning neither at the initial of Amazighe names.
- Complex Morphological System: It is a fact that the Amazighe language is agglutinative having a rather complex and rich derivational and inflectional morphology. Names can have several inflected and derived forms; a simple elimination of suffixes is not enough to reunite words families. Indeed, affixes can alter the meaning of a word.
  Similarly to other natural languages, Amazighe presents uncertainties in grammatical classes. Actually the same form is suitable for numerous grammatical categories, depending on the context in the sentence. For example, "illi" (transliterated in a french-style) can be considered as an accomplished positive verb, it means "there is" or as the name of kinship "my daughter."
- Spelling Variants: The Amazighe language has remained essentially an oral language for a long time. Therefore, the Amazighe text does not respect the standard writing convention. Furthermore, Amazighe text contains a large number of transliterated and translated NEs. These translated and transliterated words may be spelled differently and still refer to the same word with the same meaning, producing a many-to-one ambiguity. Fig. 1 shows some examples.

Amazighe examples	English translation	Entity type
(ⵎⵓⵃⴰⵎⵎⴰⴷ,muhmmad),(ⵎⵓⵃⴰⵎⵎⴰ,muhmmd),(ⵎⵃⵎⵎⴷ, mhmmd)	Mohammed	Person
(ⴰⴱⵓⴷⴰⴱⵉ, abudabi), (ⴰⴱⵓ ⴷⴰⴱⵉ, abudabi)	Abou Dabi	Location

Fig. 1. Examples of Variations in Amazighe Texts

- Lack of Linguistic Resources: We lead study on the Amazighe language resources and NLP tools (e.g., corpora, gazetteers, POS taggers, etc.). This led us to wrap up that there is a limitation in the number of available Amazighe linguistic resources in comparison with other languages. Many of those available are not relevant for Amazighe NER tasks due to the absence of NEs

annotations in the data collection. Amazighe gazetteers are rare as well and limited in size. Therefore, we tend to build our Amazighe linguistic resources in order to train and evaluate Amazighe NER systems.

3 Amazighe NER System Architecture

In this paper, we develop an hybrid architecture that is normally better than the rule-based or machine-learning systems individually. Figure 2 illustrates the architecture of the hybrid NER system for Amazighe.

The system consists of two modes: rule-based and ML-based Amazighe NER modes. The processing goes through three main phases: 1) The rule-based NER phase, 2) The feature selection and extraction, and 3) the ML-based NER phase.

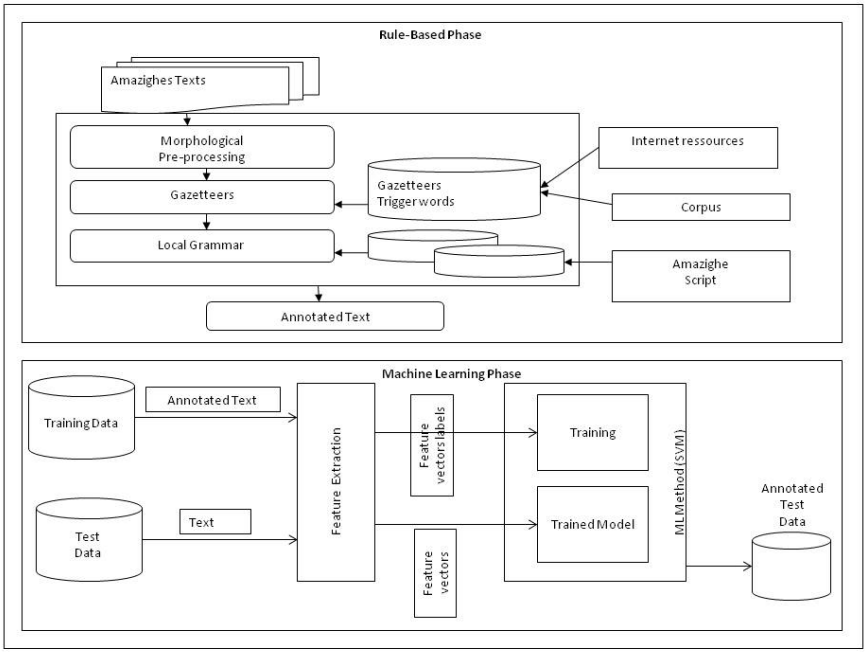

Fig. 2. Structure of our NER System

3.1 The Rule Based Phase

The rule-based component in our hybrid system is a reproduction of the NERAM system[15] using GATE[22] framework. The rule-based mode is developped with

the abilility of recognizing the 5 NEs. The recognition process used contains two principal steps: a lookup procedure, called Gazetteers, including lists of known named entities; and a finite state transducer, called Grammar, based on a set of grammar rules derived by analyzing the local lexical context relieved from our corpus (examples is provided in Figure 3). We arrive at these resources after examining several sample news articles and try to make their coverage as high as possible.

```
Rule:   TitlePerson
Priority: 30
(TITLE)
(
 {Token.kind== word}
):person
-->
{
gate.AnnotationSet person = (gate.AnnotationSet)bindings.get("person");
gate.Annotation personAnn = (gate.Annotation)person.iterator().next();
gate.FeatureMap features = Factory.newFeatureMap();
features.put("kind", "personName");
features.put("rule", "TitlePerson");
outputAS.add(person.firstNode(), person.lastNode(), "Person",
features);
}
```

Fig. 3. Example rule for Person name recognition

This rule would be able to recognize a person name based on the trigger words. Example shown in Fig. 4 would be recognized by the previous rule.

ⵎⴰⵙⵙ ⴰⵃⵎⴰⴷ ⵛⴰⵀⵉⵏ, mass Ahmad Chahin, Mr. Ahmed Chahin

Fig. 4. Example Person Name Preceded by Person Title

The GATE environment is used to build the rule-based mode. Table 1 illustrates the number of gazetteers and rules implemented within each NE type. The system contains a total of 75 rules and 24 gazetteers.

3.2 Machine Learning Phase

The ML-based phase consists on two principal steps: feature extraction and selection of ML classifiers. The first step is the feature extraction which requires the selection of classification features. The features explored are divided into various categories:

Table 1. The Number of Gazetteers and Rules in each NE Type

Named Entity Type	Rules	Gazetteers entries
Person	16	2482
Location	15	2017
Organization	13	504
Date/Time	23	170
Numerical Expressions	8	152

Context words: These are the preceding and following words surrounding the current token, ie, these are the word set adjacent to NE. This feature accounts the different contexts in which NEs appear in the training data. All of these context relations and similar information can be collected as some useful features for predicting the unknown named entities. In our implementation, features are weighted according to their distance from the current instance annotation. In other words, features which are further removed from the current instance annotation are given reduced importance.

Gazetteers: This is the gazetteer feature, gathered from the look-up gazetteers: handcrafted lists of names of person names, locations (Countries, cities, ...), organization names (association, institutes, ...), date (hours, days, years, ...) and numerical expressions (numbers, percent, ...). This feature can be determined by finding a match in the gazetteer of the corresponding named entity type.

Mention: We prepared our corpus with annotations providing class information as well as the features to be used. Actually in GATE each class has its own annotation type (Date, Person, Organization, etc.), but the Machine Learning processing resource in GATE expects the class to be a feature value, not an annotation type. So, we have created a class information in the form of a single annotation type, named "Mention", which contains a feature "class".

The second step concerns the ML classifier used in the training, testing and prediction phases. The SVM ML technique has been chosen for their high performance in NER in general and Amazighe NER in particular.

In this work, GATE, an efficient workbench that support a large number of ML algorithms, is employed as the environment of the ML task.

4 Experimental Datasets

Amazighe Language suffer from the scarcity of language technological advancements. For NER in Amazighe language, suitable corpora have until recently been unavailable, thus we have created our own corpora, besides as we mentioned in previous works we have developed a stop word list, a triggers word list, and gazetteer component, that could be more helpful for our task. In this part we introduce our resources built for Amazighe.

4.1 Corpus and Sets used

Our aim was to set up a resource comparable to more traditional general corpus used for other languages, containing a wide range of text types and topics. We have built a large corpus of Amazighe language constructed by crawling the "MapAmazighe"[21] website, which is the Amazighe information portal of "Maghreb Arab Press(MAP)", as well it is one of the largest freely available linguistic resources for Amazighe. The corpus contains more than 173 480 tokens. The corpus is actually a collection of 867 articles. Our goal was to construct a relatively heterogeneous topics, we have collected the whole news on royal activities of His Majesty King Mohammed VI (395 articles) and princely ones (93 articles), Regional (31 articles), Economics (58 articles), Social (60), Politics news (61), Sport (61), world activities (52 articles) and some general news (56 articles). We have decomposed our corpus into 4 sets, in order to minimize application execution times during the experiments. The sets "1, 2, 3, 4" respectively contain around "4168, 5273, 4963, 4281" distinct tokens. We manually annotate these data sets, using GATE that we used for this purpose, with MUC style named entity tags.

4.2 Evaluation data sets

We provide below statistical information regarding the evaluation data sets.

Set 1. The manual annotation lead us to a total of 6338 named entities. The annotated entities encompass 924 person, 1678 location, 332 organization names along with 582 date and 2822 numerical expressions.

Set 2. We preprocess this data set and the resulting set contains a total of 6827 named entities where 1452 of them are person names, 434 organization names, 1582 location names, 517 date and 2842 of them are numerical expressions.

Set 3. The manual annotation process results in the annotation of 6447 named entities with 1573 person, 1435 location, 287 organization names in addition to 744 date and 2408 numerical expressions.

Set 4. Similar to the previous data set, we obtained a total of 5039 named entities after annotation, with 936 person, 985 location, 416 organization names, 491 temporal expressions and 2211 numerical expressions.

5 Evaluation results and Analysis

In this section, we report the details of experimental setup, datasets of experiments and the evaluation results.

5.1 Metrics

In this work, we choose recall, precision and f-measure as three set-based measures . The definitions of recall, precision and F-Measure are given below:

$$Recall = \left(\frac{Correct + 0.5 * Partial}{Correct + Missing + 0.5 * Partial} \right) \quad (1)$$

$$Precision = \left(\frac{Correct + 0.5 * Partial}{Correct + Spurious + 0.5 * Partial} \right) \quad (2)$$

$$F - Measure = \left(\frac{2 * Recall * Precision}{Recall + Precision} \right) \quad (3)$$

In the preceding formulae:

- Correct corresponds to the number of named entities extracted by the system which are exactly the same as their counterparts in the answer key.
- Spurious represents the number of entities spuriously (erroneously) extracted by the system, they do not have corresponding annotations in the answer key.
- Missing is the number of named entities which are not annotated, hence missed, by the system although they are annotated in the answer key.
- Partial denotes the number of named entities extracted by the system which have corresponding entities annotated in the answer key with the same type, hence their type is correct but the tokens they contain are not exactly the same since either some tokens are erroneously missed or included by the system.

From the definitions, while recall tries to increase the number of tagged entries as much as possible, precision tries to increase the number of correctly tagged entries, and F-measure is the harmonic mean of recall and precision.

5.2 Results Obtained

The evaluation results of our system on these data sets are provided in table 2 using the above metrics. Results show that the rule-based approach leads to

Table 2. Performance of Our Rule-based System

Named Entity Type	Precision (%)	Recall (%)	F-Measure (%)
Person	98	100	99
Location	99	100	99
Organization	99	100	93
Date/Time	96	98	97
Numerical Expressions	71	87	79

good results. Apparently, Rule-based approach has best accuracy on categories of people, organization and localization as types of NE, but there are many discrepancies with the rest, this is due to the confusion that our system makes between Temporal and Numerical Expressions.

Table 3. Performance of Our System

Named Entity Recognition System	Precision (%)	Recall (%)	F-Measure (%)
Rule Based Approach	90	97	93
Hybrid (ML + Rule Based)	81	67	73

For the second experiment, we applied our hybrid system on our corpus, we splitted the corpus into training and test data, to truly know how well a machine learner is performing, for training we have selected 3 sets and 1 set for test phase.

Just to remain, we used the LibSVM SVM implementation. In this experiment, we used the linear kernel with the cost C as 0.7 and the cache memory as 100M. Additionally we used uneven margins, with τ as 0.4. The classification type is set as 'one-vs-others', meaning that the Machine Learning API will convert the multi-class classification problem into a series of binary classification problems using the one against others approach.

If we focus on results in table 3, we can easily deduce that our hybrid approach performed quite poorly in terms of precision, recall and f-measure, probably due to the nature of the dataset, distribution of our training and data sets, limited surrounding context, spelling mistakes, machine learning parameters and features used for this experiment and this clearly shows the necessity of determining appropriate feature set for the problem. Although it achieved good accuracy and we are currently working on expanding rules, testing more features to help in improving performance.

To summarize, all of the proposed systems achieve promising results on the test data set which is a meaningful contribution to NER research on Amazighe Texts, as related work is quite lacking compared to studies on other languages such as English, French, Chinese, etc., but to the best of our knowledge, our proposed system is the first to apply hybrid approach to NER on Amazighe texts.

Yet, we expect that the results should be verified on larger test corpora and can be improved by increasing the annotated training data set. Other crucial future task is to make a deeper elaboration of the employed parameters and features set to better evaluate their effects.

6 Conclusion & future works

Applying Named Entity Recognition for Amazighe language is a challenging, emerging research area, gaining more significance every day, especially due to the

increase in the size of Amazighe texts that need to be processed, but nonetheless, building a NER system for Amazighe Language is still an open problem because it exhibits characteristics different from English. In this paper, Our hybrid NER system has the ability to enrich its lexical resources with those that it learns from annotated texts through learning approach. Both the hybrid system and its rule based predecessor are evaluated on 4 data sets of different genres: news on royal activities and princely ones, financial and social news texts, regional and politic news, sport and world activities texts and some general news. These data sets are manually annotated by the authors due to the lack of available annotated corpora for NER research in Amazighe language. The evaluation results shown that our proposed method achieves promising results, but the rule based approach still perform better than our hybrid approach.

Finally, this paper envisions possible improvements on the approach in order to further increase the score ot the proposed system, including larger annotated corpus, integrating POS tagging processing, deep analysis on features set (e.g. morphological features, etc) doing to experiment with varying the configuration file to see if we can produce varied results and appliying other machine learning mode to decide which one has the best performance on our data.

References

1. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes, vol. 30, no 1, p. 3–26. (2007)
2. Tan, J. K., Benbasat, I.: The Effectiveness of Graphical Presentation for Information Extraction: A Cumulative Experimental Approach*. Decision Sciences, vol. 24, no 1, p. 167–191. (1993)
3. Costantino, M., Morgan, R. G., Collingham, R. J., Carigliano, R.: Natural language processing and information extraction: Qualitative analysis of financial news articles. In Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997, pp. 116–122. IEEE (1997)
4. Feifan, L., Jun, Z., Bibo, L., Hao, Y., Yingju, X.: Study on Product Named Entity Recognition for Business Information Extraction. Journal of Chinese Information Processing, vol. 20, No. 1, pp. 7–13. (2006)
5. Leaman, R., Gonzalez, G.: BANNER: an executable survey of advances in biomedical named entity recognition. In Pacific Symposium on Biocomputing, vol. 13, pp. 652–663. (2008)
6. Mandl, T., Womser-Hacker, C.: The effect of named entities on effectiveness in cross-language information retrieval evaluation. In: Proceedings of the 2005 ACM Symposium on Applied Computing (SAC 2005), pp. 1059–1064. (2005)
7. Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., Goranov, M.: Semantic annotation, indexing, and retrieval. In The Semantic Web-ISWC 2003, pp. 484–499. Springer Berlin Heidelberg (2003)
8. Cimiano, P.: Ontology learning from text. pp. 19–34. Springer US (2006)
9. Jin, W., Ho, H. H., Srihari, R. K.: OpinionMiner: a novel machine learning system for web opinion mining and extraction. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1195–1204. ACM (2009)

10. Nobata, C., Sekine, S., Isahara, H., Grishman, R.: Summarization system integrated with named entity tagging and IE pattern discovery. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002). Spain (2002)
11. Pizzato, L.A., Molla, D., Paris, C.: Pseudo relevance feedback using named entities for question answering. In: Proceedings of the 2006 Australian Language Technology Workshop (ALTW-2006), pp. 89–90. (2006)
12. Babych, B., Hartley, A.: Improving machine translation quality with automatic named entity recognition. In: Proceedings of EAMT/EACL 2003 Workshop on MT and Other Language Technology Tools, pp. 1–8. (2003)
13. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: 16th international conference on computational linguistics, pp. 466–471. COLING (1996)
14. 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)
15. 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)
16. 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)
17. Chaker, S.: Textes en linguistique berbère - introduction au domaine berbère. éditions du CNRS. pp. 232–242. (1984)
18. Cohen, M.: Langues chamito-sÃĬmitiques. Edouard Champion, (1924)
19. Institut Royale de la Culture Amazighe, `http://www.ircam.ma`
20. Boukhris, F., Boumalk, A., Elmoujahid, E., Souifi, H.: La nouvelle grammaire de l'amazighe. IRCAM, Rabat (2008)
21. Amazighe Information Portal of "Maghreb Arab Press (MAP)", `http://www.mapamazighe.ma`
22. General Architecture for Text Engineering, `https://gate.ac.uk/`