Journal of Theoretical and Applied Information Technology
15th March 2023. Vol.101. No 5
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TOWARDS A FULLY CLOUD-BASED PLATFORM FOR ARABIC NLP

MOHAMMED NASRI[1], SARA CHAOUKI[2], MOSTAFA SAADI[3] AND NOREDINE GHERABI[4]

[1,2,3,4]LaSTI Laboratory, University Sultan Moulay Slimane, ENSA Khouribga, Béni Amir, ENSA, B.P 77, Khouribga, Morocco.

E-mail: [1]mohammed.nasri@gmail.com, [2]srh.chaouki@gmail.com, [3]saadi_mo@yahoo.fr, [4]n.gherabi@usms.ma

ABSTRACT

Cloud Computing is getting more and more debated in the IT industry today. Its evolution is leading the next generation of internet services. Natural Language Processing (NLP) is a field of Artificial Intelligence that focuses on understanding, manipulating and generating human language by machines. Thus, the NLP is really at the interface between computer science and linguistics. It, therefore, concerns the ability of the machine to interact directly with humans. Arabic NLP is very poor compared to other languages such as English or German due to the complexity of this language and the lack of resources. In this work, we propose a new system for Arabic NLP based entirely on the Cloud. This system is based on two steps. It firstly uses a bridge between Arabic and other developed languages (English for occurrence) and then uses of the already developed features for that language. Hence those features apply not on the Arabic text but instead on the translated (the English) one. In some cases, the result needs to be in Arabic, in which case, we use the bridge another time to translate English result into Arabic. This can either be used in real NLP systems, such as Translation, IR, QA, Sentiment Analysis, or for validation or comparison purposes, especially for those who work in NLP and use other approaches. Experiments have been performed on a prototype we developed and the results obtained are satisfactory for this first version.

Keywords: *Arabic NLP, Cloud-based platform, SaaS NLP*

1. INTRODUCTION

Natural Language Processing or simply NLP is an area of Artificial Intelligence that aims to make machines understand or generate Human languages. This research field at the crossroads of linguistics, semantics and machine learning has reached a sufficient qualitative level to leave the laboratories and invest the productive economy ([1], [2] and [3]). Indeed, companies in many sectors (IT, Customer Relationship, Telecommunications, etc.) of activity are now using NLP to enhance their data and increase their productivity. Frequent use cases are indexing and categorizing content, data mining, consumer feedback analysis and information monitoring.

With the Cloud Computing and especially Software As A Service (SaaS) offers, new ideas for NLP have become increasingly less complicated. Indeed, there is no longer need to care about software licenses, complex installations or user machine performance. Moreover, SaaS offers cover several NLP features such as entity recognition, sentiment analysis or summarization. There is no need for particular expertise in this area, all one needs is providing the content to the service and get the result. This is true for some languages such as English or Spanish, but not general for the other languages such Arabic.

Recall that Arabic is a Semitic language with various dialects based on diversified geographical areas. According to Ethnologue 2019[1], there are 273,989,700 Arabic speakers around the world. It is also the official language for almost 26 countries and is officially written and spoken by their governments [4]. It is spoken ordinarily on the radio, used in official speeches, courses and university conferences. Despite all this, Arabic NLP (ANLP)

[1] http://www.ethnologue.com/statistics/size

remains less developed, due to many factors, among them the complex morphology of Arabic, the luck of resources in ANLP and the variety of its dialects.

In this work, we propose a new approach in ANLP field based entirely on the Cloud Computing, which combines the features provided by the four giant providers of the Cloud (Amazon, Microsoft, Google, IBM), connects them and processes Arabic language. Our objective is not to provide or implement a specific NLP feature, but rather, an idea for a new general-purpose platform based on the Cloud.

The paper is organized as follows; The next section discusses some related works. Section 3 provides an overview of the existing Cloud services related to NLP. Our approach as well as an implementation and experiments are detailed in Section 4. The paper ends with the conclusion and some perspectives.

2. RELATED WORK

Using SaaS services for NLP is not a new idea, SaaS-based works have already been introduced in sentiment analysis [5], information extraction [6], insights discovery [7] or voice call analytics [8]. These are examples of works that have already used the Cloud approach to solve NLP problems. their common characteristic is that they are feature-oriented, they all deal with a specific feature, be it sentiment analysis, information retrieval or other.

On the other hand, ideas as well as works for Arabic NLP platforms have already been introduced by Jaafar et al. [9], Zerrouki [10] and Sakhr Software.

Jaafar et al. [9] introduced SAFAR (Software Architecture For ARrabic) which is an open source, cross-platform and modular platform dedicated for Arabic NLP. As described by the authors, SAFAR includes: 1) Resources needed for different Arabic NLP services, 2) Basic levels modules of language, especially those of the Arabic language, namely morphology, syntax and semantics and 3) Applications for the ANLP such as Information Retrieval, Question/Answering, Named Entity Recognition, etc. SAFAR has been developed for several years by a whole team, which has been integrating several modules, and continues developing and adding custom features to its core.

Zerrouki [10] introduces Adawat, a tool that integrates many applications, APIs and corpora such as: Light stemmer, verb conjugator, morphology analyzer, Spell checker, Text to speech system, Mishkal diacrtizer, vocalized texts corpus, synonyms dictionary, collocations etc. In Adawat, Zerrouki uses mainly rule based approach to build rules and data. Adawat is a good idea for a new platform, but it is still at an early stage.

Sakhr Software, provided by Sakhr Software Company, transforms its research in NLP into industry-first commercial software and solutions. Sakhr provides solutions for Arabic, including: Machine Translation, Optical Character Recognition, Speech Technology, Knowledge Management, Advanced Research Services, Professional Translation & Localization. Sakhr software is used by both government and business entities across multiple industries, including education, financial services, media, social services, technology, and telecommunications. The only drawback is that it is not free.

3. OVERVIEW OF THE CLOUD SERVICES RELATED TO NLP

3.1 Introduction

Recall that Cloud computing enables companies to consume services on demand. These services are organized into three successive levels: the infrastructure level (IaaS), the platform level (PaaS) and the software level (SaaS).

In this paper, we looked at the SaaS offers provided by the four giants: Amazon Comprehend, Microsoft Azure Text Analytics, Google Cloud Natural Language and IBM Watson Natural Language Understanding.

It is important to know what languages are processed and what features are considered, which is interesting for several reasons, among them:

1. To know the dominant language for the features of each cloud provider.
2. To know the less dominant languages and whose features still need to be developed.

3.2 Amazon Comprehend

Amazon Comprehend[2] provides the following features: Dominant language, Entity recognition,

[2] https://aws.amazon.com/comprehend/

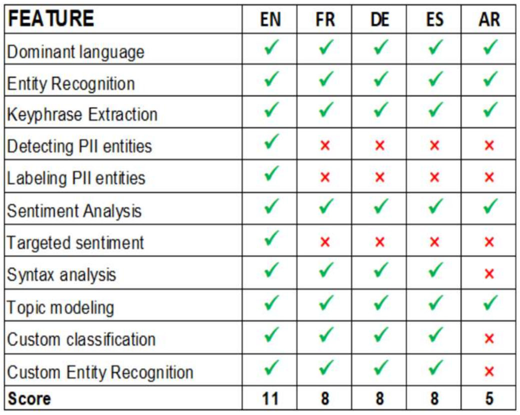
Keyphrase Extraction, Detecting PII (Personally Identifiable Information) entities, Labeling PII entities, Syntax analysis, Targeted sentiment, Syntax Analysis, Topic modelling, Custom Classification, and Custom Entity Recognition. Amazon Comprehend has been used in several real applications, such as Insight Discovering [7], Market Demand Analysis [11] or Entity Anchoring [12].

Besides, these features have been developed by Amazon for several languages. As first work, we classified Amazon features in table alongside with the corresponding languages (features in rows, languages in columns and cells with or without the check mark that means this feature is or is not provided for this language)[3] . This helps perform some analysis such as identify the most/less developed languages and the more/less developed feature. For such table to be more useful, we added a score to each language (at the bottom of each column) by counting the number of the features that are developed for that language.

We noticed that English has the score of 11, which means that it is the most dominant/developed language, followed by French, German, Spanish and Portuguese in second rank, whereas, the less dominant languages were Bengali, Russian, Chinese and Arabic.

The first version of this table listed many columns, and for simplicity, clarity and especially focus reasons, we only concentrated on the four languages: English (EN), French (FR), German (DE), Spanish (ES) in addition to Arabic (AR). Table 1 represents this short version.

Table 1: Features Provided By Amazon Comprehend



FEATURE	EN	FR	DE	ES	AR
Dominant language	✓	✓	✓	✓	✓
Entity Recognition	✓	✓	✓	✓	✓
Keyphrase Extraction	✓	✓	✓	✓	✓
Detecting PII entities	✓	×	×	×	×
Labeling PII entities	✓	×	×	×	×
Sentiment Analysis	✓	✓	✓	✓	✓
Targeted sentiment	✓	×	×	×	×
Syntax analysis	✓	✓	✓	✓	×
Topic modeling	✓	✓	✓	✓	✓
Custom classification	✓	✓	✓	✓	×
Custom Entity Recognition	✓	✓	✓	✓	×
Score	11	8	8	8	5

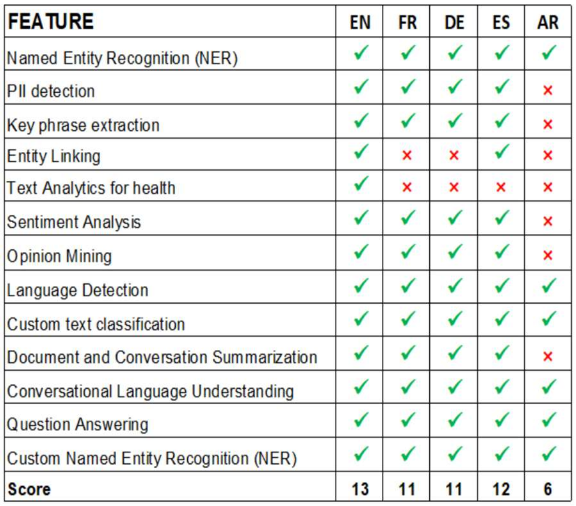
In the rest of this paper, we will only consider those languages. The first four are chosen thanks to their dominance and development (they have the highest score), where Arabic is only for illustration purposes, any other non-developed or less dominant language could be used instead. We have some preferences for Arabic since it's our native language and we aim to develop it more and more.

If we zoom in, we can point out that the features: Dominant language, Entity Recognition, Keyphrase Extraction and Sentiment analysis are provided for all languages. However, Detecting PII entities, Labelling PII entities, Targeted sentiment analysis, Syntax analysis, Topic modelling, Custom classification and Custom Entity Recognition are still to develop.

3.3 Azure Text Analytics

Azure Text Analytics[4] provides the following features: Named Entity recognition (NER), Personally Identifiable Information (PII) detection, Key phrase extraction, Entity linking, Text Analytics for health, Sentiment analysis, Opinion mining, Language detection, Custom text classification, Document and Conversation Summarization, Conversational language understanding, Question Answering, Custom named entity recognition (NER). Azure Text Analytics has been used in Big Data [13], Mood detection [14] or Customer feedback Enhancement [15].

Table 2: Features Provided By Azure Text Analytics



FEATURE	EN	FR	DE	ES	AR
Named Entity Recognition (NER)	✓	✓	✓	✓	✓
PII detection	✓	✓	✓	✓	×
Key phrase extraction	✓	✓	✓	✓	×
Entity Linking	✓	×	×	✓	×
Text Analytics for health	✓	×	×	×	×
Sentiment Analysis	✓	✓	✓	✓	×
Opinion Mining	✓	✓	✓	✓	×
Language Detection	✓	✓	✓	✓	✓
Custom text classification	✓	✓	✓	✓	✓
Document and Conversation Summarization	✓	✓	✓	✓	×
Conversational Language Understanding	✓	✓	✓	✓	✓
Question Answering	✓	✓	✓	✓	✓
Custom Named Entity Recognition (NER)	✓	✓	✓	✓	✓
Score	13	11	11	12	6

[3] This table is not shown in this paper due to its size.

[4] https://azure.microsoft.com/en-us/products/cognitive-services/text-analytics/#overview

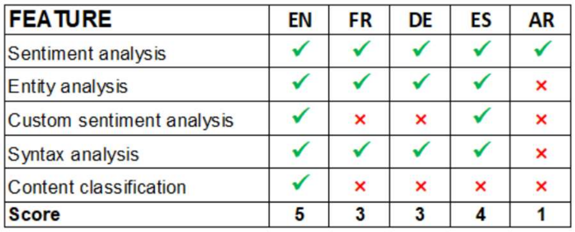
In the same way as Amazon Comprehend, we classified the features of Azure Text Analytics in table 2 with the corresponding languages and scores.

According to the results mentioned in table 2, we notice, here too, that English is the most dominant language, followed by French, German, and Spanish, whereas, Arabic remains the less dominant.

3.4 Google Natural Language AI

Google Natural Language AI[5] provides less features than Amazon Comprehend or Azure Text Analytics, here are features offered by Google Natural Language AI: Sentiment analysis, Entity analysis, Custom Sentiment analysis, Syntax analysis and Content classification. Table 3 presents the summary of these features per language.

Table 3: Features provided by Google Natural Language AI



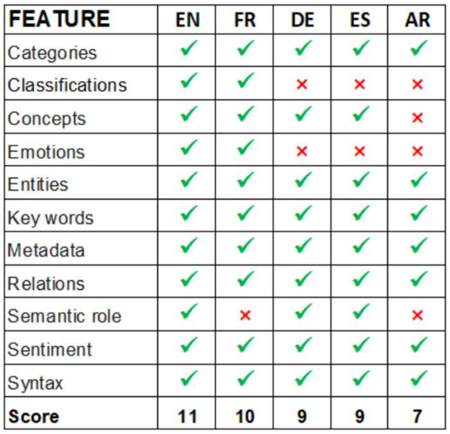
FEATURE	EN	FR	DE	ES	AR
Sentiment analysis	✓	✓	✓	✓	✓
Entity analysis	✓	✓	✓	✓	×
Custom sentiment analysis	✓	×	×	✓	×
Syntax analysis	✓	✓	✓	✓	×
Content classification	✓	×	×	×	×
Score	**5**	**3**	**3**	**4**	**1**

As usual, we notice from the table 3 that English remains the most dominant language, followed by Spanish, French and then German, while Arabic is the least dominant. We also highlight that the following features: sentiment analysis, entity analysis, custom sentiment analysis, and content classification are not available in all languages.

3.5 IBM Watson Natural Language Understanding

IBM Watson Natural Language Understanding[6] provides the following features: Categories, Classifications, Concepts, Emotions, Entities, key words, Meta-data, Relations, Semantic role, Sentiment and Syntax. Table 4 presents the summary of these features per language.

Table 4: Features provided by IBM Watson Natural Language Understanding



FEATURE	EN	FR	DE	ES	AR
Categories	✓	✓	✓	✓	✓
Classifications	✓	✓	×	×	×
Concepts	✓	✓	✓	✓	×
Emotions	✓	✓	×	×	×
Entities	✓	✓	✓	✓	✓
Key words	✓	✓	✓	✓	✓
Metadata	✓	✓	✓	✓	✓
Relations	✓	✓	✓	✓	✓
Semantic role	✓	×	✓	✓	×
Sentiment	✓	✓	✓	✓	✓
Syntax	✓	✓	✓	✓	✓
Score	**11**	**10**	**9**	**9**	**7**

Table 4 shows that English is the most dominant language, followed by French, German and Spanish, while Arabic is the least dominant. We also highlight that the following features: Classifications, Concepts, Emotions and Semantic role are not available in all languages.

3.6 Analysis and Comparison

With the purpose of analyzing, comparing and making conclusion, we gathered in a single summary table the four previous tables 1, 2, 3 and 4 respectively related to the four NLP Cloud modules: Amazon Comprehend, Azure Text Analytics, Google Cloud Natural Language and IBM Watson Natural Language Understanding. This summary table helps us:

1. Easily compare the features of each NLP module,
2. Extract the most dominant languages and the least dominant ones,
3. Highlight the features that are provided by the four providers for each language,
4. Find out which features are the least developed, and
5. Identify the features that are not developed for Arabic but are already developed for the other languages (English, French, German, etc.).

Once we started analyzing, we figured out that the four providers don't give the same name to the same (or equivalent) feature. In fact, we found that many features that do the same thing and that they have the

[5] https://cloud.google.com/natural-language

[6] https://www.ibm.com/cloud/watson-natural-language-understanding

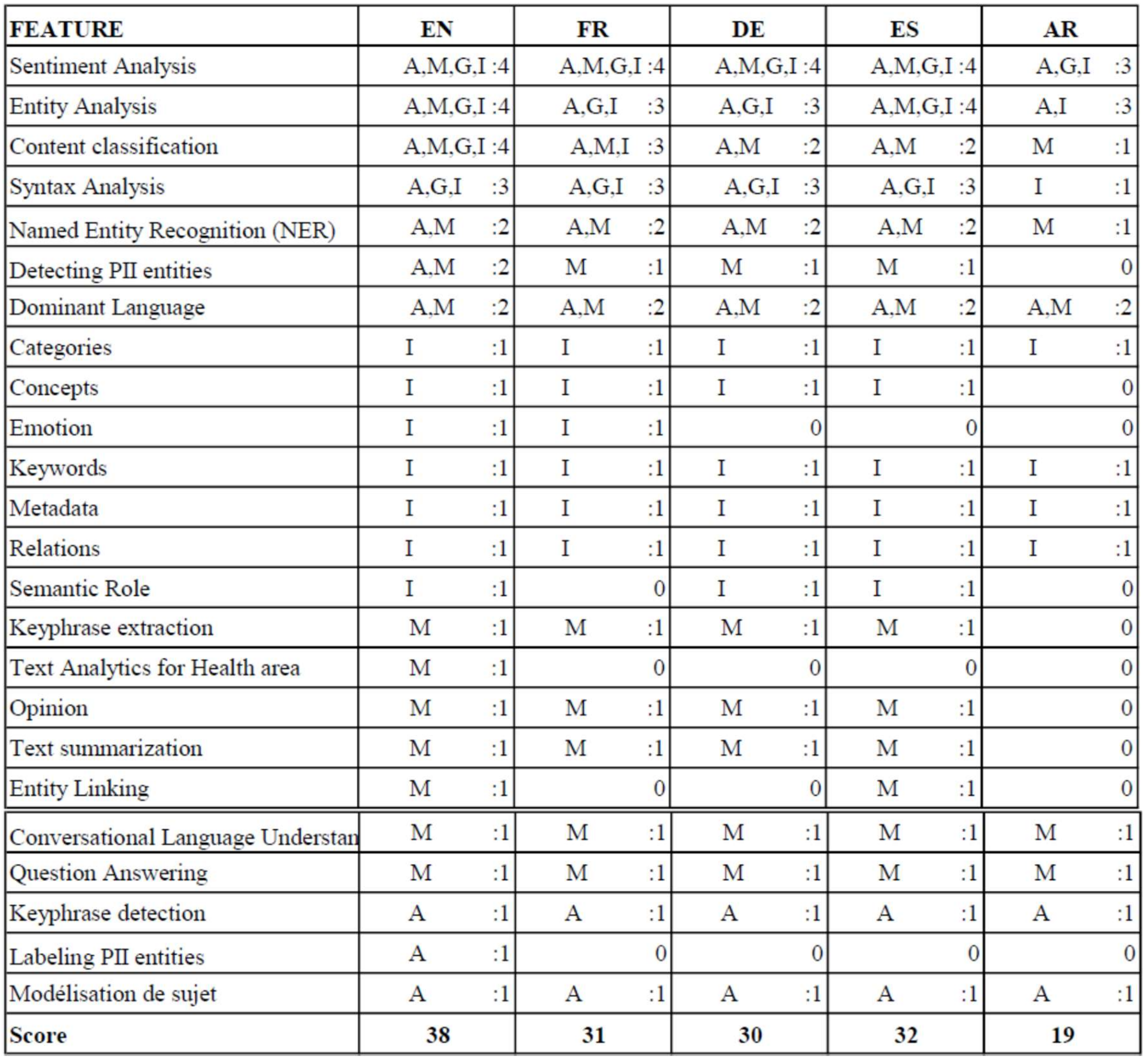
same description, their names were different depending on the provider. For instance, the “Custom Classification” feature provided by Amazon Comprehend is named by Google Cloud as “Content Classification”, by Microsoft Azure as “Custom Text Classification” and by IBM as “Classifications”. A second example is the feature “Language Detection” provided by Azure which is named “Dominant Language” on Amazon Comprehend.

The second work we performed was, therefore, to standardize the names of the features. So, we choose a unique name for similar features, for example: a) Amazon Custom Classification, b) Azure Custom Text Classification, c) Google Content Classification and d) Classifications of IBM have been named in “Content Classification”. Hence, we renamed most of the features.

A new classification has been performed in this stage, taking into account the new names, the five languages and the four providers (Amazon, Microsoft, Google and IBM are respectively referred to as A, M, G, I).

We finally gave a score of 0 to 4 to each feature, depending on whether this feature is not provided by anyone (score = 0), by only one provider (score = 1), by two (score = 2), by three (score = 3), or by all of them (score = 4). The global score of each language was computed by adding the scores of each feature. This is very important to know the degree of maturity of this language regarding the features and the providers. Table 5 gives details about this summary.

Table 5: Summary



FEATURE	EN	FR	DE	ES	AR
Sentiment Analysis	A,M,G,I :4	A,M,G,I :4	A,M,G,I :4	A,M,G,I :4	A,G,I :3
Entity Analysis	A,M,G,I :4	A,G,I :3	A,G,I :3	A,M,G,I :4	A,I :3
Content classification	A,M,G,I :4	A,M,I :3	A,M :2	A,M :2	M :1
Syntax Analysis	A,G,I :3	A,G,I :3	A,G,I :3	A,G,I :3	I :1
Named Entity Recognition (NER)	A,M :2	A,M :2	A,M :2	A,M :2	M :1
Detecting PII entities	A,M :2	M :1	M :1	M :1	0
Dominant Language	A,M :2	A,M :2	A,M :2	A,M :2	A,M :2
Categories	I :1	I :1	I :1	I :1	I :1
Concepts	I :1	I :1	I :1	I :1	0
Emotion	I :1	I :1	0	0	0
Keywords	I :1	I :1	I :1	I :1	I :1
Metadata	I :1	I :1	I :1	I :1	I :1
Relations	I :1	I :1	I :1	I :1	I :1
Semantic Role	I :1	0	I :1	I :1	0
Keyphrase extraction	M :1	M :1	M :1	M :1	0
Text Analytics for Health area	M :1	0	0	0	0
Opinion	M :1	M :1	M :1	M :1	0
Text summarization	M :1	M :1	M :1	M :1	0
Entity Linking	M :1	0	0	M :1	0
Conversational Language Understan	M :1	M :1	M :1	M :1	M :1
Question Answering	M :1	M :1	M :1	M :1	M :1
Keyphrase detection	A :1	A :1	A :1	A :1	A :1
Labeling PII entities	A :1	0	0	0	0
Modélisation de sujet	A :1	A :1	A :1	A :1	A :1
Score	**38**	**31**	**30**	**32**	**19**

According to the scores on table 5, we see that English, Spanish and French are the most developed. We also notice that the English column does not have any 0 in its cells, which means that all the features of NLP are already developed for that language. However, in the columns of the French, German and Spanish languages, we notice that there are features which are not yet developed (cells with the score 0). The situation is worse for Arabic, whose column contains many features that are not developed yet.

Fortunately, the translation feature for Arabic is provided by the four providers from and to English, Spanish, French and German. We can then use a bridge between Arabic and a target language (one of the four) and use the features available in that targeted language. For example, we can translate an Arabic text to English, apply Sentiment Analysis on that English text (which is equivalent) and return the result as if Sentiment Analysis was initially applied on the original Arabic text itself. In general, this approach is appliable to any poor language (Arabic, Bengali, Russian, etc.) whenever a bridge could be built to a developed language (English, French, Spanish, etc.). In our case, we choose Arabic as poor language and English as developed one.

From table 5, we highlight the features that are not provided for Arabic, but that are provided for English, for example 1) Detecting PII entities, 2) Concepts, 3) Emotion, 4) Semantic role or 5) Text summarization, and we will apply them to Arabic.

4. OUR SYSTEM

4.1 Overview

In our system, we propose to create a new system/platform for Arabic NLP entirely based on the Cloud. This system uses a bridge between Arabic and other developed languages and exploits their already developed features. An overview of this process is illustrated in figure 1 and detailed in below.

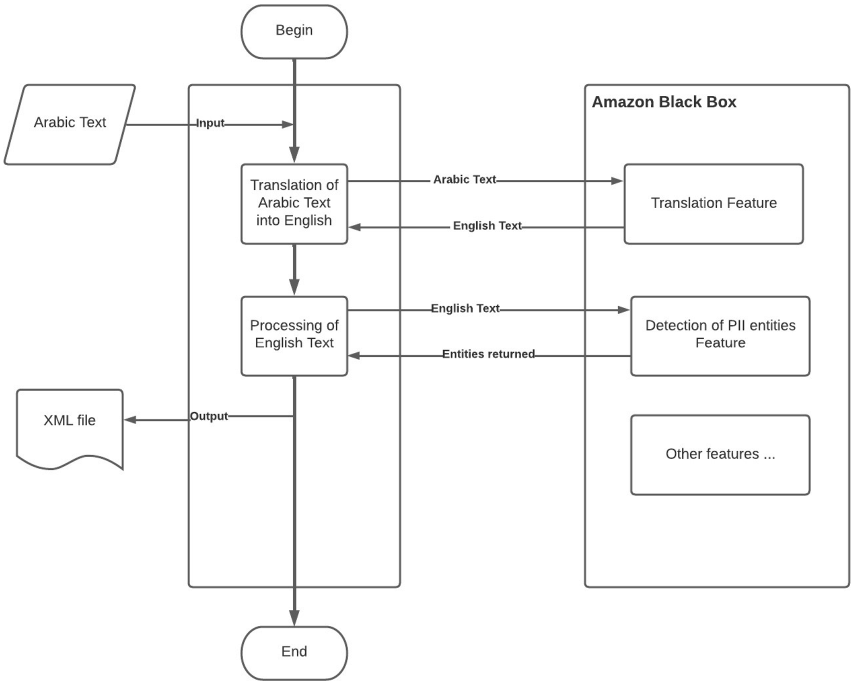


Figure 1: Overview of our process

First of all, for each feature that is not provided in Arabic but provided for other languages, we 1) use the bridge to that language, so that the Arabic text is translated to its equivalent in English, this text will then 2) be analyzed by the feature that is available in English and 3) the result is finally returned to the user.

A fourth step may be necessary. In fact, for systems like Information Retrieval or Question Answering that requires the result to be returned to user in Arabic, we use the bridge another time to translate that result from English to Arabic. This step is not always necessary since features like sentiment analysis whose result is either positive, negative or neutral does not require any translation.

4.2 Implementation and Experiments

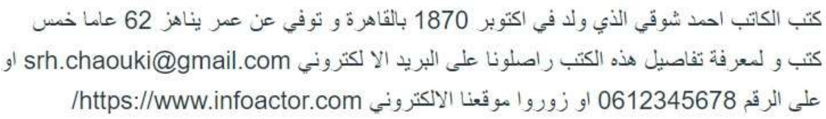
As first implementation, we decide to use Amazon provider and implement the feature Detecting PII entities. These choices are not strategic, but remain random choices we made in order to give a concrete example.

We then used, in a Java project, the API provided by Amazon for its module Comprehend[7]. We implemented the process illustrated in figure 1 and performed some tests to ensure that everything works as expected.

To test this first version, we downloaded the Khaleej Dataset[8], a subset of SANAD [16]. Khaleej dataset is a large collection of Arabic news articles that can be used in different Arabic NLP tasks such as text classification. This dataset is made of 45,500 articles in 7 categories (Culture, Finance, Medical, Politics, Religion, Sports, Tech), each category contains 6,500 files.

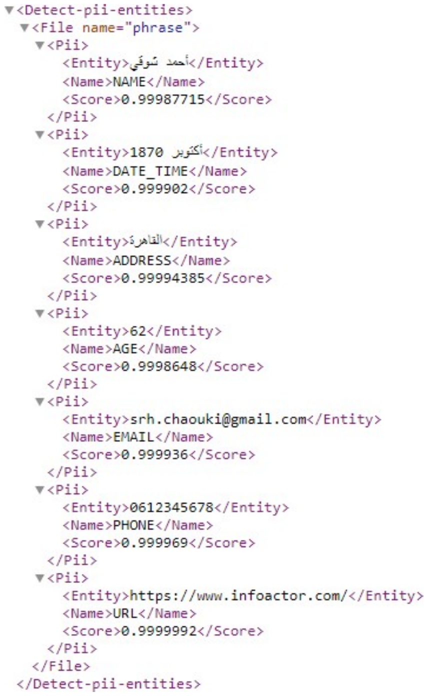
We then ran our program on the corpus. For each article, our program translated the content into English, and try to execute the selected feature. The execution of our program took 2 seconds per 10 files (due to the remote call of the API) with no reported error or exception (for the whole corpus). The results have been put in XML files, one article per file. Afterward, we manually analyzed 5 files (taken randomly), the results were reliable.

Let’s illustrate a whole example of this process. Figure 2 shows an Arabic text (extracted from the corpus) to be processed by our system. Figure 3 shows the result of this analysis.



كتب الكاتب احمد شوقي الذي ولد في اكتوبر 1870 بالقاهرة و توفي عن عمر يناهز 62 عاما خمس
كتب و لمعرفة تفاصيل هذه الكتب راصلونا على البريد الا لكتروني srh.chaouki@gmail.com او
على الرقم 0612345678 او زوروا موقعنا الالكتروني https://www.infoactor.com/

Figure 2: Example of Arabic text extracted from AlKhaleej dataset



```
▼<Detect-pii-entities>
  ▼<File name="phrase">
    ▼<Pii>
       <Entity>أحمد شوقي</Entity>
       <Name>NAME</Name>
       <Score>0.99987715</Score>
     </Pii>
    ▼<Pii>
       <Entity>1870 أكتوبر</Entity>
       <Name>DATE_TIME</Name>
       <Score>0.999902</Score>
     </Pii>
    ▼<Pii>
       <Entity>القاهرة</Entity>
       <Name>ADDRESS</Name>
       <Score>0.99994385</Score>
     </Pii>
    ▼<Pii>
       <Entity>62</Entity>
       <Name>AGE</Name>
       <Score>0.9998648</Score>
     </Pii>
    ▼<Pii>
       <Entity>srh.chaouki@gmail.com</Entity>
       <Name>EMAIL</Name>
       <Score>0.999936</Score>
     </Pii>
    ▼<Pii>
       <Entity>0612345678</Entity>
       <Name>PHONE</Name>
       <Score>0.999969</Score>
     </Pii>
    ▼<Pii>
       <Entity>https://www.infoactor.com/</Entity>
       <Name>URL</Name>
       <Score>0.9999992</Score>
     </Pii>
   </File>
 </Detect-pii-entities>
```

Figure 3: The result obtained after execution by our system

[7] https://docs.aws.amazon.com/comprehend/latest/dg/using-the-api.html

[8] https://www.kaggle.com/datasets/haithemhermessi/sanad-dataset

It is clearly seen in figure 3 that the system has identified a name, a date, an address, an age, an email, a telephone number and a URL. Although this simply uses services already provided, it provides a new approach to Arabic NLP that could help many researchers in many situations.

Based on these experiments and the results obtained, we can conclude that our approach is reliable and that the steps proposed to create a bridge between Arabic and English have proven their effectiveness for the chosen functionality in particular, and for the others in general.

4.3 End Word

Last but not least, if we look deeper into the process we used, we notice that we firstly use a bridge (based on Amazon features), and then explore one of the features from Amazon. This was of our choice, we could continue using the bridge of Amazon but with the features of one of the other providers (IBM, Azure, or Google) which could lead to better results. In future works, we will, for the same features:

1. Combine the 4 bridges to the features provided by the 4 providers,
2. Analyze and compare results, and
3. Finally, integrate into the platform the better combination.

5. CONCLUSION AND PERSPECTIVES

In our work, we proposed a prototype for Arabic NLP platform fully-based on Cloud. This system uses a bridge between Arabic and other developed languages and exploits their already developed features.

To prove the feasibility of the concept, we developed a prototype based on Amazon as provider and English as language. We used a bridge between Arabic and English (Translation Service provided by Amazon), the text input in Arabic is translated into English by Amazon. The result (English text) is then performed via one of the features provided by Amazon for the English language, in our case, we used for example the feature “Personal Identifiable Information Detection”.

Our choice for Amazon as Cloud provider, English as language and Detecting the Personally Identifiable Information entities as feature are nothing but illustrations. The same principles could be applied to any other Cloud provider, any other language or feature.

This same principle could be applied to other features, which leads to a so-called fully cloud-based platform. Such platform could be beneficial in many situations, for instance: It can obviously be used in real NLP systems, such as Translation, Information Retrieval, Question Answering, Sentiment Analysis, etc. It can also be used for validation or comparison purposes, for those who work in NLP and use other approaches. Indeed, it is a common concern for researchers in experiment stage to be able to compare their results with other platforms, other tools, regardless of the way these latter are developed, as long as they provide results for the same corpus/corpora.

We could finally affirm that the goal has been achieved by realizing the prototype for a new system for Arabic NLP based entirely on the Cloud.

As perspectives, we can go on implementing the other features provided by the four providers for English and that are not provided yet for Arabic.

REFERENCES:

[1] Sann, Raksmey, and Pei-Chun Lai. “Understanding homophily of service failure within the hotel guest cycle: Applying NLP-aspect-based sentiment analysis to the hospitality industry.” International Journal of Hospitality Management 91 (2020): 102678.

[2] Khan, Shahnawaz, and Mustafa Raza Rabbani. “Artificial intelligence and NLP-based chatbot for islamic banking and finance.” International Journal of Information Retrieval Research (IJIRR) 11.3 (2021): 65-77.

[3] Tallamraju, Ravindra Babu. “Geographical Address Models in the Indian e-Commerce.” Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.

[4] Chohan, Muhammad Nadeem, et al. “PHONEMIC COMPARISON OF ARABIC AND ENGLISH.” Hamdard Islamicus 43.1&2 (2020): 387-405.

[5] Satyanarayana, G., J. Bhuvana, and M. Balamurugan. “Sentimental Analysis on voice using Amazon Comprehend.” 2020 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2020.

[6] Dias, Dulan S., Madhushi D. Welikala, and Naomal GJ Dias. “Identifying racist social media comments in sinhala language using text analytics models with machine learning.” 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, 2018.

[7] Mohanta, Raj Kumar. “Discover Insights and Relationships in Text Using Amazon Comprehend.”

[8] Sudarsan, V., and Govind Kumar. “Voice call analytics using natural language processing.” Int. J. Stat. Appl. Math 4 (2019): 133-136.

[9] Jaafar, Y., and K. Bouzoubaa. “SAFAR: software architecture for Arabic language processing.” (2017).

[10] Zerrouki, Taha. “Towards an open platform for arabic language processing.” (2020).

[11] Sabbir Hossain, Md, Nishat Nayla, and Annajiat Alim Rasel. “Product Market Demand Analysis Using NLP in Banglish Text with Sentiment Analysis and Named Entity Recognition.” arXiv e-prints (2022): arXiv2204.

[12] DeYoung, Jay, et al. “Entity anchored ICD coding.” arXiv preprint arXiv:2208.07444 (2022).

[13] Janani, R., and S. Vijayarani. “Text Analytics in Big Data Environments.” Big Data Applications in Industry 4.0. Auerbach Publications, 2022. 125-143.

[14] Ferdiana, Ridi, Wiiliam Fajar Dicka, and Faturahman Yudanto. “MOOD DETECTION BASED ON LAST SONG LISTENED ON SPOTIFY.” ASEAN Engineering Journal 12.3 (2022): 123-127.

[15] Sormunen, Hanna-Maria. “Enhancing Customer feedback processing with Machine Learning in Microsoft Azure.” (2022).

[16] Einea, Omar, Ashraf Elnagar, and Ridhwan Al Debsi. "Sanad: Single-label arabic news articles dataset for automatic text categorization." Data in brief 25 (2019): 104076.