**University of Ahmed DRAIA -ADRAR-**

**Department of Mathematics and Computer Science**

A project presented for the

**Master's degree in Computer Science**

# Converting Dialectal Arabic to Modern Standard Arabic

**Elwannas HIRI**

Supervised By

**Mr. Mohamed Amine CHERAGUI**

**June 2021**

الجمهورية الجزائرية الديمقراطية الشعبية

République algérienne populaire et démocratique

وزارة التعليم العالي والبحث العلمي

Ministère de l'enseignement supérieur et de la recherche scientifique

UNIVERSITE AHMED DRAYA - ADRAR
BIBLIOTHÈQUE CENTRALE
Service de recherche bibliographique
N°.......................B.C/S.R.B//U.A/2021

جامعة احمد دراية - ادرار
المكتبــــة المركــزية
مصلحة البحث البيبليوغرافي
الرقم.............م م/م.ب.ب /ج.أ/2021

# شهادة الترخيص بالإيداع

انا الأستاذ(ة): شراقيا محمد أمين

المشرف مذكرة الماستر.

الموسومة بـــ : Converting Dialect ARABIC TO Modern Standard ARABIC

من إنجاز الطالب(ة): الدناس هيري

و الطالب(ة):

كلية : العلوم و التكنولوجيا

القسم : الرياضيات و الاعلام الآلي

التخصص: أنظمة ذكية

تاريخ تقييم / مناقشة: 2021/06/26

أشهد ان الطلبة قد قاموا بالتعديلات والتصحيحات المطلوبة من طرف لجنة التقييم / المناقشة، وان المطابقة بين النسخة الورقية والإلكترونية استوفت جميع شروطها.

وبإمكانهم إيداع النسخ الورقية (02) والاليكترونية (PDF).

- امضاء المشرف:

شراقي محمد أمين

ادرار في : 06 JUIL. 2021

مساعد رئيس القسم:

مساعد رئيس قسم الرياضيات والإعلام الآلي
مكلف بالتدريس والتعليم في التدرج
مازوزي حاج

وزارة التعليم العالي والبحث العلمي ★ جامعة أدرار ★ كلية العلوم والتكنولوجيا
قسم الرياضيات والإعلام الآلي

ملاحظة :لاتقبل أي شهادة بدون التوقيع والمصادقة.

## Abstract

The field of studying the arabic dialects has attracted a lot of researchers recently, with regards to the importance of this area in many domains of the time and the rising demand for this kind of needs. In this project we present our AMSAC corpus (Algerian dialect Modern Standard Arabic Corpora), a collection of more than 14k sentences, the largest corpus for Algerian dialect to our knowledge. We also present our model LAHDJA, a translation model for the Algerian dialect to the MSA. LAHDJA has achieved the best results compared to the Meftouhe model with a 15.13 BLUE score.

**Key words:** Dialect, translation model, corpus

## الملخّص

استقطب مجال دراسة اللّهجات العربيّة الكثير من الباحثين في الآونة الأخيرة، لما له من أهميّة في العديد من مجالات العصر وكذا الطّلب المتزايد على هذا النّوع من الاحتياجات. في هذا المشروع، نقدّم مدوّنة الجمل أمساك، وهي مدوّنة لأكثر من 14 ألف جملة، والتي تعتبر أكبر مدوّنة جمل للّهجة الجزائرية على حد علمنا. نقدّم أيضًا نموذجنا لهجة، وهو نموذج ترجمة للّغة الجزائرية إلى اللُّغَةِ العَرَبِيَّة الفُصحَى. حقّق لهجة أفضل النّتائج مقارنة بنموذج مفتوح بنتيجة: 15.13

**الكلمات المفتاحيّة:** نموذج ترجمة، لهجة

# Dedication

To my parents, to my brother Abbes and his little man Hude, to my sister Assia, to my sister Nawel and everyone who contributed to this work. I dedicated this work.

*Elwannas*

# Acknowledgements

All the Thank and the Respect to our Supervisor Mr. Mohamed Amine CHERAGUI for Guiding us in every step of this project. A very special thanks to Dr. Mohammed MEDIANI and Dr. Abdelghani DAHOU which they helped me during all steps in my thesis. To anyone interested in studying the Arabic language. Thanks to the Algerian community generally and 1001Night-Tech specially, namely Romaissa NEDDJAR. To all my friends which they are my team namely Mohammed Cherif ABBOU, Amine ABDEDAIM, Shaimaa BENAICHAOUI, Yassine BENLARIA, Adel DAFI, Wassen ELDJANABI, Salem GUERROUT, Dounia KHECHAB, Youness NASSAMOU and Molay Ali YOUSFAT thank you for supporting me until last seconds.

*Elwannas*

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| AD | Algerian Dialect |
| BnT | BLEU score without tokenization |
| DA | Dialectal Arabic |
| EGY | Egypt |
| GLF | Gulf |
| LEV | Levantine |
| LSTM | Long short term memory |
| MSA | Modern Standart Arabic |
| NLP | Natural language processing |
| NPS | Number of parallel sentences |
| NSPT | Number of sentences per test |
| NTRS | Number of training sentences |
| NTS | Number of test sentences |
| NVS | Number of validation sentences |
| OSV | Object Subject Verb |
| OVS | Object Verb Subject |
| PPL | Perplexity |
| SOV | Subject Object Verb |
| SVO | Subject Verb Object |
| TD | Tnusian Dialect |
| UAE | United Arabic Emerat |
| VOS | Verb Object Subject |
| VSO | Verb Subject Object |

# Introduction

The natural language processing is a research field situated at the intersection of several disciplines: Artificial Intelligence, Theoretical Computer, Statistical, Linguistics, ...etc. Its main objective is the design and the development of software able to automatically process linguistic data, i.e. data expressed in a natural language (whether Standard or Dialect). In recent decades, natural language processing has seen a real ascension, which has allowed us to move out of the standard language ("academic and well-structured language") to the dialect, which remains a heterogeneous version of the standard language closer to humans. Scientifically and also socio-economically, several specialized companies and products have been created. Today, we speak of automatic spelling correction, automatic summarizing, natural language database interrogation, sentiment analysis, etc. But the main topic of this field remains Machine Translation. If in the past, having a text translated by a machine was utopian, the technological developments and the constant efforts of many researchers in Machine Translation (MT) make it possible today. MT is the process of translating a text from a source language to a target language by a machine without any human intervention. For machine translation from or to Arabic, a lot of work has been done, but in terms of performance they are still a fair way behind other languages such as English or French. The objective of our project is to contribute to the development of the Arabic language processing by developing an environment for the machine translation from the Algerian dialect to the standard Arabic language using the deep learning approach. This project also allows us, in addition to our translation system "LAHDJA", to build an important linguistic resource, which is the corpus "AMSAC". In order to achieve our objective, we have structured our thesis in 04 chapters, which are:

- Introduction to Machine Translation
- Challenges for Arabic Machine translation
- The conception of our translation model "LAHDJA"
- Implementation and results

# Chapter 1

# Introduction to Machine Translation

## 1.1 Machine Translation History

The table below shows a brief history about MT [3],[5],[6],[7],[9],[11],[13],[14], [15],[18],[26].

| Generation | Period | Name | Updates {Events, Changes} |
|---|---|---|---|
| 1 | 1948 - 1960 | The begining | - Warren Weaver[1] suggested in his letter to use computer for translation |
| | | | - Yehoshua Bar-Hillel[2], created a translator at IBM which translates more than 60 sentences. He predicted than MT would not be an issue in 3-5 years. |
| | | | - Victor Yngve[3], the first, in 1954, brought out the first who dealt with MT. |
| 2 | 1960 - 1966 | Parsing and disillusionment | - Earlier in 1960s, parsing and disilluionment was the only field of reseach in MT. |
| | | | - Computational linguistique born, thanks to David G. Hays[4] |
| | | | - First international conference for MT. |
| | | | - In 1964, ALPAC[5] Studies the chance of MT. |
| | | | - ALPAC declared that MT is waste of time considering time consuming. |
| 3 | 1966 - 1980 | New birth and hope | - 1970 Start of the project REVERSO by a group of Russian researchers. |
| | | | - Creation of WEATHER system by Alai Colmerauer |
| | | | - 1978 Creation of ATLAS[6] by FUJITSU[7] for Korean-Japanese translation. |
| 4 | 1980 - 1990 | Japanese invaders | - In 1983, NEC[8] creates a translation method based on an algorithm called PIVOT. |
| | | | - A rule-based system called PENSEE from OKI[9] was released in 1986. |
| | | | - Based on rules, Hitachi[10] created its own translation system. |
| 5 | 1990 - now | Web and new vague of translators | - In 1993, The theme of C-STAR[11] project was MT. |
| | | | - 2005, Translation websites come out. |
| | | | - 2010 28% of intenet users used MT. 50% planned to do. |

Table 1.1: Machine Translation History

## 1.2 Machine Translation Approaches

Since the first MT system, NLP researchers generally and MT Scientist specifically developed different approaches to automate the translation. Deep Neural Networks architectures were changing for image recognition [8] and speech recognition[10]. [28] MTs, benefit from the existed data and boost the performance. It costs less time and money comparing to Rule-based or SMT systems. The table below present MT approaches[7].

| | Paradigms | Concept | Approaches |
|---|---|---|---|
| 1 | Rule-based Machine Translation Figure (1.2) | Based on language theory. Language experts spend time to extract rules for a specific language in order to create system to generate target language translation | - Direct Translation, |
| | | | - Transfer Based |
| | | | - Interlingua |
| 2 | Data-driven Machine Translation Figure (1.3) | Based on examples, learn from existed sentences to generate new states | -Statistical Machine Translation |
| | | | - Neural Machine Translation[28] |
| | | | Example-based Machine Translation |
| 3 | Hybrid Machine Translation | Hybrid machine translation is a single-based framework that incorporates many machine translation strategies. | Rule-based Approach + Data-driven approach |

Table 1.2: Machine Translation Approaches

[1] Warren Weaver was a US scientist, a mathematician and an administrator of science (July 17, 1894 – November 24, 1978

[2] Yehoshua Bar-Hillel was a mathematician, philosopher, and linguist from Israel. He was a forerunner in machine translation and formal linguistics. (https://tinyurl.com/t1g3r01-PFE-YehoshuaBarHillel)

[3] Victor Y. was a professor of linguistics at the University of Chicago (1925–2012).He was an early adopter of computer linguistics and natural language processing. (https://tinyurl.com/t1g3r01-PFE-Victor-Yngve)

[4] David Glenn Hays (November 27, 1928 – July 26,1995), was a computational linguist and social scientist known for early work in machine translation(https://tinyurl.com/t1g3r01-PFE-David-G-Hays)

[5] Automatic Language Processing Advisory Committee

[6] Current version is 14

[7] Fujitsu Limited is a Japanese multinational information and communications technology https://tinyurl.com/t1g3r01-PFE-FUJITSU

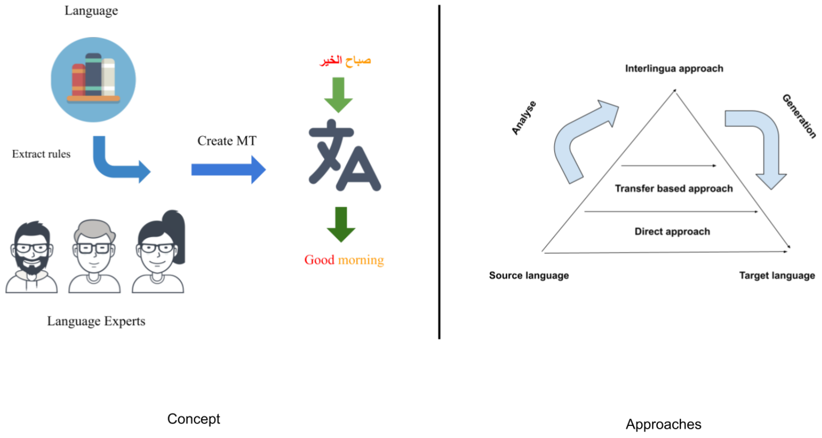
[8] NEC Corporation is a Japanese IT and electronics multinational company based in Minato, Tokyo. (https://tinyurl.com/t1g3r01-PFE-NEC)

[9] Founded in 1881 OKI Electric Industry Co, is a Japanese manufacturer of telecommunications

[10] Hitachi: is a Japanese multinational conglomerate company headquartered in Chiyoda, Tokyo, Japan. https://tinyurl.com/t1g3r01-PFE-HITACHI

[11] Consortium for Speech Translation Advanced Research

### 1.2.1 Rule Based (Visual view of concept & approach)

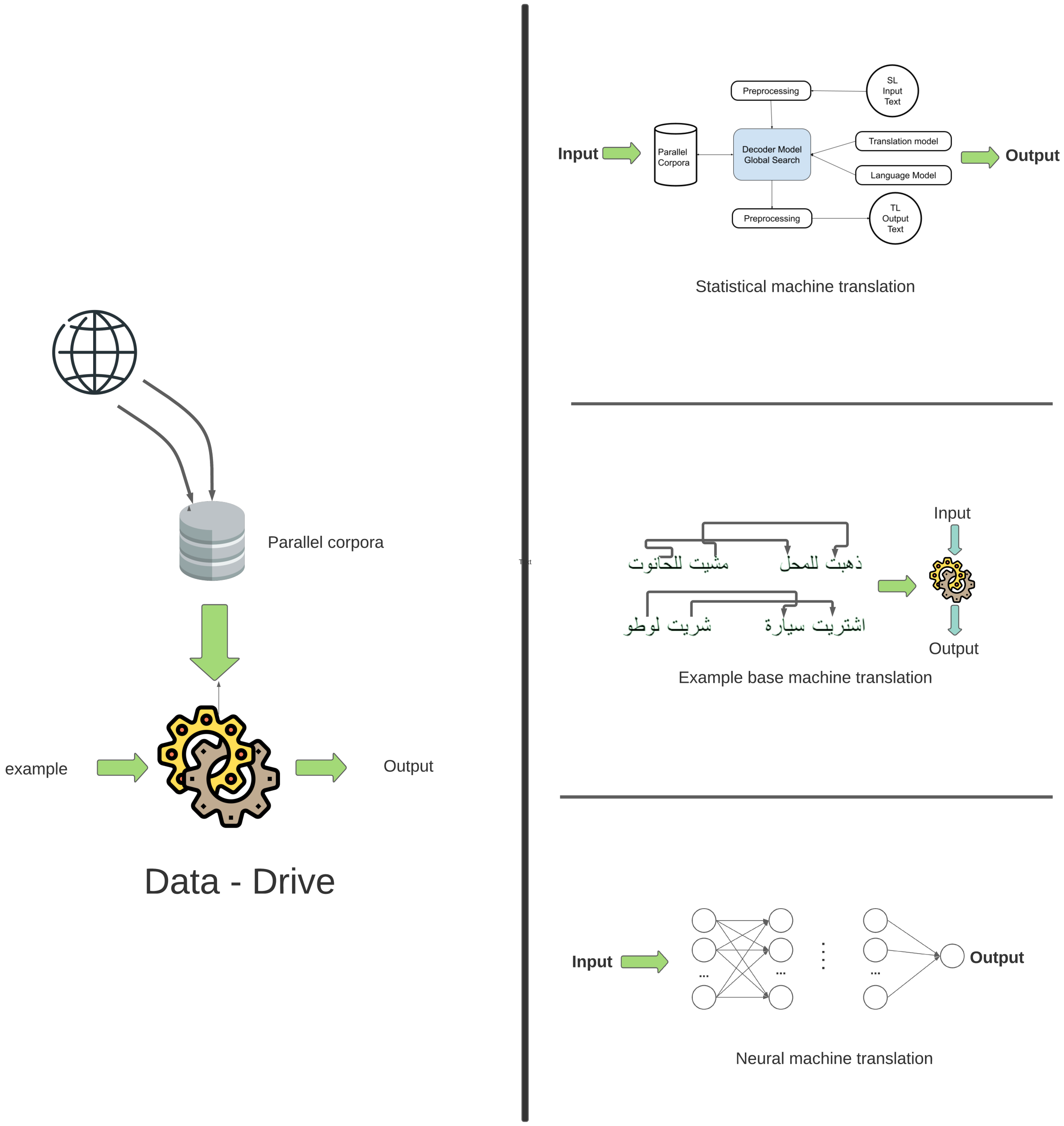


Concept

Approaches

Figure 1.1: Rule Based (Visual view of concept & approach)

### 1.2.2 Data-driven (Visual view of concept & approach)



Statistical machine translation

Example base machine translation

Neural machine translation

Figure 1.2: Data Driven(Visual view of concept & approach)

### 1.2.3 Hybrid MT (Visual view of concept & approach)

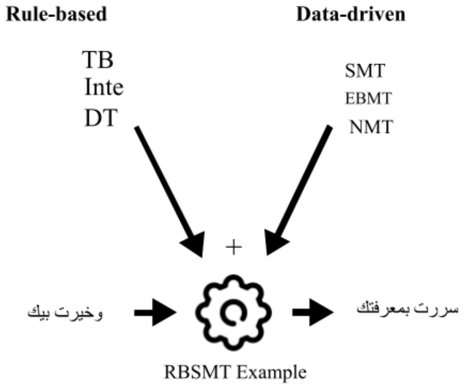


Figure 1.3: Hybrid Machine Translation (Visual view of concept & approach)

## 1.3 Challenges for Arabic Machine Translation

The Arabic language has more than 12,000,000 words[12], which is a huge number compared to other languages like English 600,000 words and French 150,000 words. Arabic differs in many variations. Starting from letters, Arabic has 29 letters. Furthermore, ”ض ”(dhad), ”ظ” (tha'a) and ”غ” (ghain) are special letters. The structure of the Arabic sentence is flexible. 120 Arabic pattern morphology exist[24]. Unlike other languages, the Arabic language has many challenges in the NLP domain generally and Machine Translation precisely. The table below present Arabic Machine Translation challenges [1].

| | Challenge | Problem | Example | Translated Example |
|---|---|---|---|---|
| 1 | diacritization | With diacritization, vocabulary the size will expose which will cause time, coast, performance problems for MT. Without diacritization, the vocab size will be reduced but cause a problem which is a limited context. | word كتب | |
| | | | كَتَبَ | He wrote |
| | | | كُتُبٌ | Books |
| | | | كُتِبَ | It is written |
| | | | Sentences كتب الدرس | |
| | | | كُتِبَ الدَّرْسُ | Written lesson |
| | | | كَتَبَ الدَّرْسَ | He wrote the lesson |
| 2 | Word ambiguity | A word which has multiple meanings | خال | Empty,<br>battalion, imagined |
| 3 | Letter ambiguity | A letter that connected to a word which has multiple meanings | بالقلم | ***Using*** a pen |
| | | | بالسيارة | ***By*** the car |
| | | | بالباب | ***At*** the door |
| 4 | Agglutination word | A complex word which can represent a sentence | ويتكلمون | And they are talking |
| | | | و | And |
| | | | ي | Are |
| | | | تكلم | Talking |
| | | | ون | They |
| 5 | Diagrams and trigrams | ”Words” are not considered as words in Arabic language | في | In |
| | | | على | On |
| | | | من | From |
| | | | إلى | To |
| | | | عن | About |
| 6 | Sentence syntax | Flexible sentence syntax<br>VSO[13]<br>SVO[14]<br>OVS[15] (less used) | Mohamed went to school | |
| | | | ذهب محمد الى المدرسة | Went Mohammed to school |
| | | | محمد ذهب الى المدرسة | Mohamed went to school |
| | | | الى المدرسة ذهب محمد | To school went Mohamed |

Table 1.3: Challenges for Arabic Machine Translation

[12] https://tinyurl.com/t1g3r01-PFE-Arabic-wn

[13] Verb Subject Object

[14] SVO: Subject Verb Object

[15] OVS: Object Verb Subject

# Chapter 2

# Translation from Arabic Dialect to MSA (State of the Art)

## 2.1 Introduction

Typically, the Arabic is the official language of the Arab world used in the official domains. But for daily communication and non official talks, the non-standard language (dialect) is used. In fact, being the dialect that is primarily used, involves a wide interest in this area in the NLP domain. In this chapter, we will present a brief review about Translation from Arabic dialect to MSA including The Difference between the Arabic dialect and MSA, why is this kind of translation important? And some previous works in this domain.

## 2.2 The difference between the Arabic dialect and MSA

Generally, the modern standard Arabic is the official language overall of the Arabic countries, it is used in the newspapers, official talks..etc, it is a modern version of the classical Arabic (CA) that used in the Quran and in the earliest literature, it has a linguistic rules and a typographic system of writing and it is standard in all Arabic countries. In parallel, the primary language used in daily talks and social networks is the Arabic dialect (non-standard form Arabic), each country has a specific dialect language, and sometimes there are more than one dialect in the same country. These dialects are usually spoken rather than being written, and they are a mix of other languages such as Berber, English, French. In dialect languages there are no set standards for writing them. Because there are no writing norms to adhere to, so it can be written with different forms that are all valid. In the table 2.1 below, we present a brief difference between the MSA and the dialect language.[16, 12]

| Features | MSA | Dialect |
|---|---|---|
| The use | News paper, formal broadcast programs, religious practice | Daily talks , non-official talks |
| Originality | Modern version of the classical Arabic | A mixed ancient local tongues and by European languages such as French, English |
| Flexibility | Standard in all the Arabic countries | Each country has its own dialect, and sometimes there is more than one dialect in the same country |
| Grammar | Linguistic rules and a typographic system of writing | No rule based and usually spoken |
| Vowels | Sensitive to the case ending (for example in the plural forms) | There is no case ending |
| Syntactic level | The Verb-Subject-Object order (VSO) and (SVO) are more used then (OVS) and (OSV) | Free word order |
| Other characteristics | A complex morphology and a rich vocabulary | Many forms are all acceptable since there are no writing rules as reference |
| Other characteristics | Msa orthography contains only Abjd-alphabetic | Some letters can be replaced by numbers. Ex: 3 replace the letter “ع” |

Table 2.1: Differences between the Arabic dialect and MSA

## 2.3 Importance of translating Dialect Arabic to Modern Standard Arabic

Several systems have been developed for the MSA translation because the NLP is more likely to focus on the standard form of any language. In parallel, dialect translations were understudied at the time. It’s just recently that it has attracted researchers due to the rising demand for them.

In adding, This sort of translation technology can be used in different domains and many purposes, such as commercial purposes. (multipurpose)

Furthermore, The fact that Arab people are more likely to use the non-standard language in dealing with the internet and social networks which are increasingly used. As shown in the table 2.2 [1], the most rated apps and used in Algeria, Tunisia and UAE are social media applications such as Messenger, Facebook, Instagram ...etc , the thing that provides a huge amount of available data that can be exploited by the researchers while developing systems in the NLP domain. (availability)

Dealing with the standard form of any language provides more flexibility and control of the context. As a result, dialect translation is critical at this time. (control)

Moreover, The field of sociology and context categorization can benefit by translating dialect materials to the standard form. So it’s easy to study society’s ideas and problems by using their social network’s expressions. (classification) [16, 12]

[1] https://www.similarweb.com/ [Last view 5/06/2021]

| Algeria | | Tunisia | | UAE | |
|---|---|---|---|---|---|
| Android | Apple | Android | Apple | Android | Apple |
| Shaareit | Messnger | Messenger | Messenger | Photography | Photo Video |
| SuperNet VPN | Facebook | Snapseed | Facebook | ALHOSN UAE | CapCut |
| Phoneix | Instagram | WhatsApp | Instagram | COVID19 UAE | ALHOSN UAE |
| Facebook lite | TikTok | SuperNet VPN | TikTok | SEHA | COVID-19 UAE |
| Tiktok | YouTube | Instagram | YouTube | DHA | DHA |
| Whatsapp | CapCut | Water Color Sort | CapCut | Catwalk Beauty | TikTok |
| Instagram | Crowd Battle 3D | Facebook Lite | Crowd Battle 3D | BOTIM l | ToonMe |
| Kwal | Google Maps | TikTok | Google Maps | Super VPN | Instagram |
| Messnger | Snapchat | Catwalk Beauty | Snapchat | TikTok | BOTIM |
| Cat Walk beauty | Truecaller | Alibaba.com | Truecaller | Instagram | Telegram Messenger |

Table 2.2: Most applications used in Algeria, Tunisia and UAE

## 2.4 Machine Translation Models(Related works, previous models)

| N | Year | Group | Paradigm | Approach | Additional Technics | Source - Target | NSPT | Corpora | Results (Bleu Scoring) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2012 | zbib [27] | Data Driven | MTurk | / | LEV-EGY | / | LDC20 06*-09*[2] | 14.05 |
| 2 | | | | | / | LEV-EGY | / | | 17.79 |
| 3 | | | | | / | EGY - LEV | / | | 11.21 |
| 4 | | | | | / | MSA - EGY | / | | 14.34 |
| 5 | | | | | / | MSA - LEV | / | | 12.29 |
| 6 | 2014 | Sadat [22] | Data Driven | Rule-based | / | TD - MSA | 50 | / | 14.32 |
| 7 | 2015 | Meftoh 15 [16] | Data Driven | Phrase based | KN/WB | **ALG - MSA** | 500 | PADIC | **15.10 / 14.64** |
| 8 | | | | | | ANB - MSA | | | 14.44 / 13.95 |
| 9 | | | | | | **MSA - ALG** | | | **13.55 / 13.05** |
| 10 | | | | | | MSA - ANB | | | 12.54 / 11.72 |
| **11** | 2018 | Meftoh 18 [17] | Data Driven | Phrase based | / | **ALG - MSA** | 500 | PADIC | **15.1** |
| 12 | | | | | / | ANB - MSA | | | 14.44 |
| 13 | | | | | / | **MSA - ALG** | | | **13.55** |
| 14 | | | | | / | MSA - ANB | | | 12.54 |
| 15 | 2018 | Baniata [2] | Neural MT | Single | Bi - LSTM | LA - MSA | 2000 | MPCA + PADIC | 0.17 |
| 16 | | | | MTLA[3] | | LA - MSA | | | 0.41 |
| 17 | | | | Single | | MA - MSA | | | 0.16 |
| 18 | | | | MTLA | | MA - MSA | | | 0.3 |
| 19 | 2020 | Sghaier [23] | Data Driven | Rule-based | / | TD - MSA | 100 | / | 55.22 |

Table 2.3: Previous work on AD - MSA Machine Translation

[2] https://catalog.ldc.upenn.edu/
[3] Multi task Learning Approach

## 2.5 Corpora

The tables below resume all the statistics of previous works PADIC[16](2.4), MADAR[4](2.6), Dial2MSA[19](2.5) and

- PADIC

| Corpora | Creation Group | N.Dialects | NPS | Dialect | Words | Vocab |
|---|---|---|---|---|---|---|
| PADIC | TORJOMAN | 5 | 6400 | **ALG** | **38707** | **8966** |
| | | | | ANB | 38428 | 9060 |
| | | | | TUN | 37259 | 10215 |
| | | | | SYR | 39286 | 9825 |
| | | | | **MSA** | **40906** | **9131** |
| | | | | PAL | 39286 | 9195 |

Table 2.4: PADIC Corpus Statistics

- Dial2MSA

| | | | | | | |
|---|---|---|---|---|---|---|
| Dial2MSA | Crowdsourcing | 4 | 5500 | EGY | 77800 | 17399 |
| | | | 5000 | MGR | 53351 | 18856 |
| | | | 6000 | LEV | / | / |
| | | | | GLF | / | / |

Table 2.5: Dial2MSA Statistics

- MADAR

| Corpora | Creation Group | N.Dialects | NPS | Dialect | Words | Vocab |
|---|---|---|---|---|---|---|
| MADAR | CAMeL | 26 | 12000 | BEI | 67216 | 28457 |
| | | | | CAR | 74517 | 27731 |
| | | | | DOH | 63663 | 26054 |
| | | | | **MSA** | **129994** | **25096** |
| | | | | RBT | 202114 | 32654 |
| | | | | TUN | 65848 | 28923 |
| | | | 2000 | ALP | 10580 | 6283 |
| | | | | ALE | 11671 | 6184 |
| | | | | **ALG** | **11643** | **6250** |
| | | | | AMM | 11746 | 6370 |
| | | | | ASW | 11999 | 6442 |
| | | | | BGH | 10772 | 6503 |
| | | | | BNG | 11568 | 6216 |
| | | | | BSR | 10308 | 6382 |
| | | | | DMS | 10637 | 6288 |
| | | | | FES | 11691 | 6623 |
| | | | | JDD | 10582 | 6150 |
| | | | | JRS | 10997 | 6096 |
| | | | | KHR | 11809 | 6297 |
| | | | | MSL | 11250 | 6540 |
| | | | | MSC | 11653 | 6606 |
| | | | | RYD | 11103 | 6252 |
| | | | | SLT | 12491 | 5945 |
| | | | | SAN | 17272 | 6397 |
| | | | | SFX | 10640 | 6105 |
| | | | | TRP | 11538 | 6164 |

Table 2.6: MADAR Corpus Statistics

## 2.6 Conclusion

Recently, translation from Arabic dialect to MSA has attracted a lot of developers and started to be studied, starting from the Middle-east dialects to the Maghreb dialects. In this chapter we have presented a review about the difference between the dialect language and the MSA, we have also discussed the importance of this kind of study in the multi purpose such as commercial and social studies and finally a brief of previous works and models related to this area of domain.

# Chapter 3

# Conception of Translation Model ”LAHDJA”

### 3.1 Introduction

The fact of being the non-standard form of the Arabic(dialect) as the officially spoken language in the daily talks and the social conversations, has given it more attention by the researchers in the ANLP domain especially in the translation area. In this field of study, a number of efforts have been adopted, the most of them focusing on the Middle-east dialects. In contrast, the Algerian dialect had less attention and it is just beginning to be studied. In this chapter we will present the conception of our proposed translation model LAHDJA, Starting by introducing our corpus AMSAC, our approach to building it, statistics and finally our approach to build the LAHDJA model.

## 3.2 General Architecture

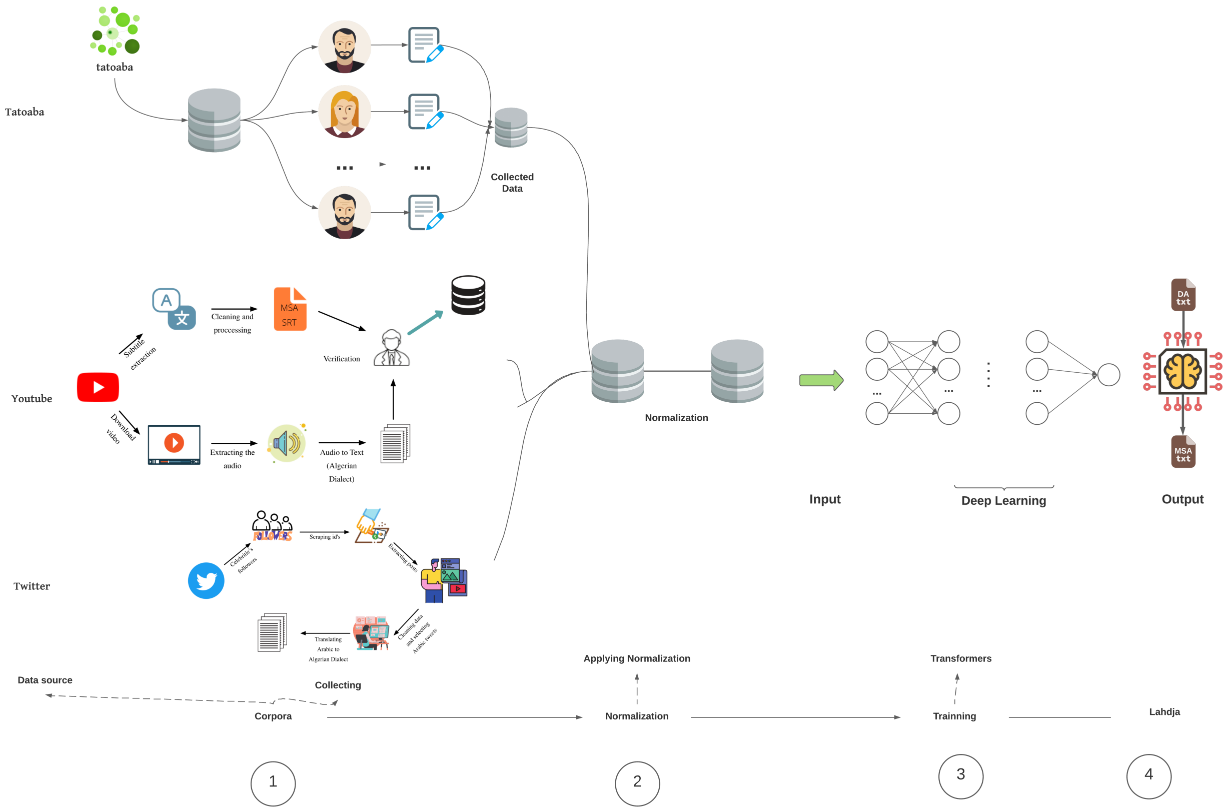


Figure 3.1: General Architecture of AMSAC + LAHDJA

## 3.3 Building of AMSAC

In order to create our AMSAC corpora (Algerian dialect Modern Standard Arabic Corpora), we have been using tree resources, which are: Tatoeba, Twitter, Youtube. Data collection was in varying proportions. In the following, we will explain our methodology for collecting data with corresponding statistics for each source.

### 3.3.1 Data source

#### Tatoeba

The website Tatoeba[1] , provides more than 160 different language data sets for machine translation. We selected the Arabic languge. The data set has been expanded to include 32,000 sentences.

[1] https://tatoeba.org/

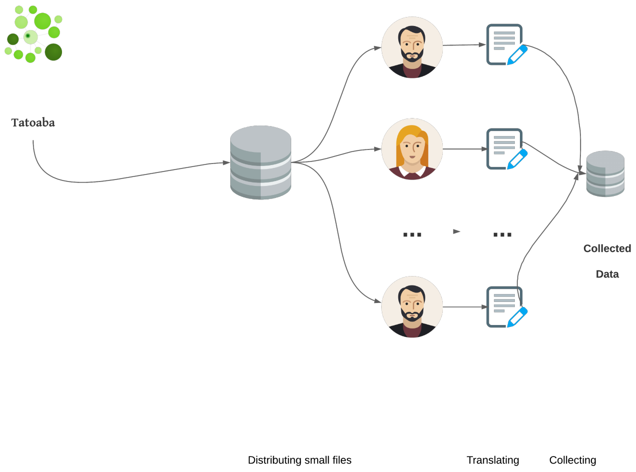


Figure 3.2: Tatoeba process, visual explanation

### Twitter

Twitter[2] is a fantastic social networking network that allows you to download tweets, list of follower from specific id. We chose two Algerian superstars[3] whose work is aimed at the Algerian people to target the Algerian tweets. Both have over 470k and 300k followers on Twitter. We started by scraping the followers with twint[4]. Next, we scraped all tweets from each public Twitter account using the same method. Finally, we got 3,715,093 tweet.

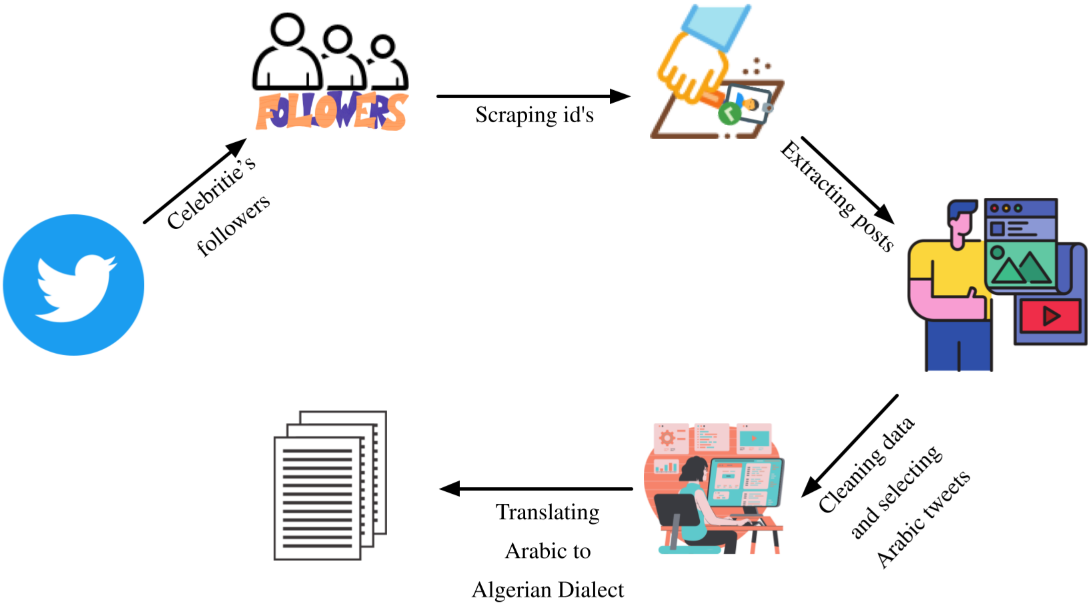

FOLLOWERS
Scraping id's
Celebritie's followers
Extracting posts
Cleaning data and selecting Arabic tweets
Translating Arabic to Algerian Dialect

Figure 3.3: Twitter process, visual explanation

[2] https://www.twitter.com

[3] 1: DZjoker , https://twitter.com/DZjokerOfficiel;
2: ZaroutaYoucef, https://twitter.com/ZaroutaYoucef

[4] https://github.com/twintproject/twint

### Youtube

Youtube has many Algerian creator which they use AD as main language in their videos.

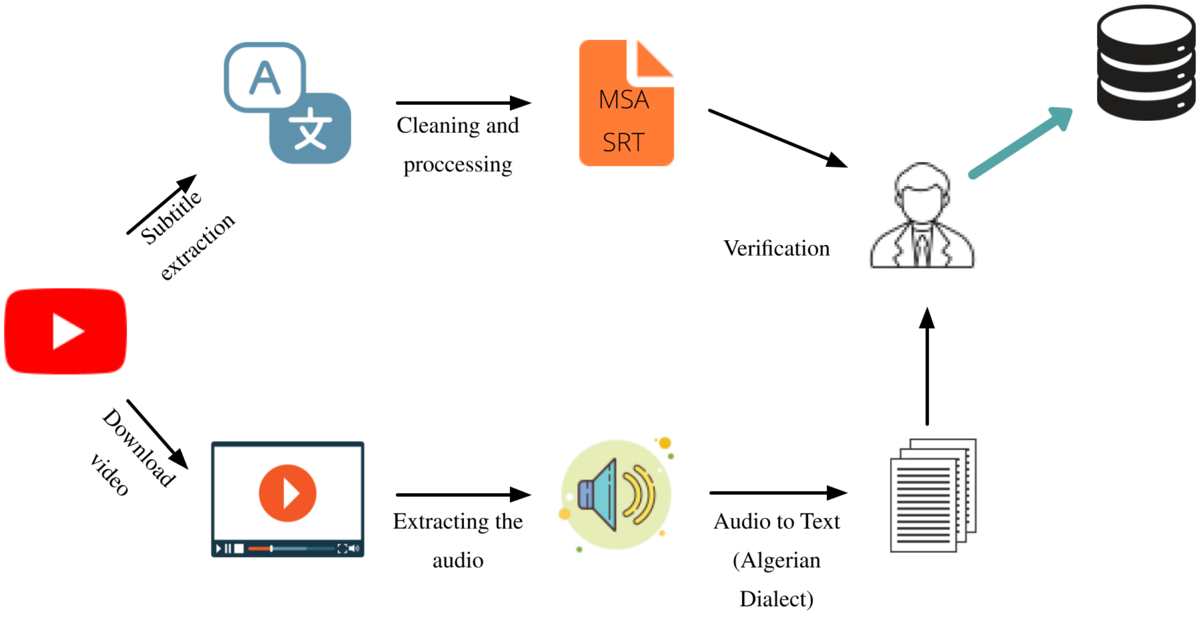


Figure 3.4: Youtube process, visual explanation

### 3.3.2 Process of building AMSAC

#### Part 1 (Tatoeba)

The database is downloaded from the site. There are 23000 phrases in the data collection. We divided it into files that are smaller. There are 100 phrases in each file. We have produced a post in which we asked the public to help us establish the company via a Facebook group titled 1001tech. We develop and transmit files depending on what users request. Some people don't have a machine. We may therefore construct editable sheets online with Google Sheets. Through addition, a group choose to just write messages in Messenger. Over 200 people reacted to our post over the first week. We received more than 116 files. After 6 weeks the post went deep on Facebook as 1001Tech posts have been published a day.

We collect non-translated files to optimize our usage for the Tatoeba data collection. We paid for the translation of the materials to four people afterward. The sum of these four translators translated phrases was 5600.

### 3.3.3 Files received by regions

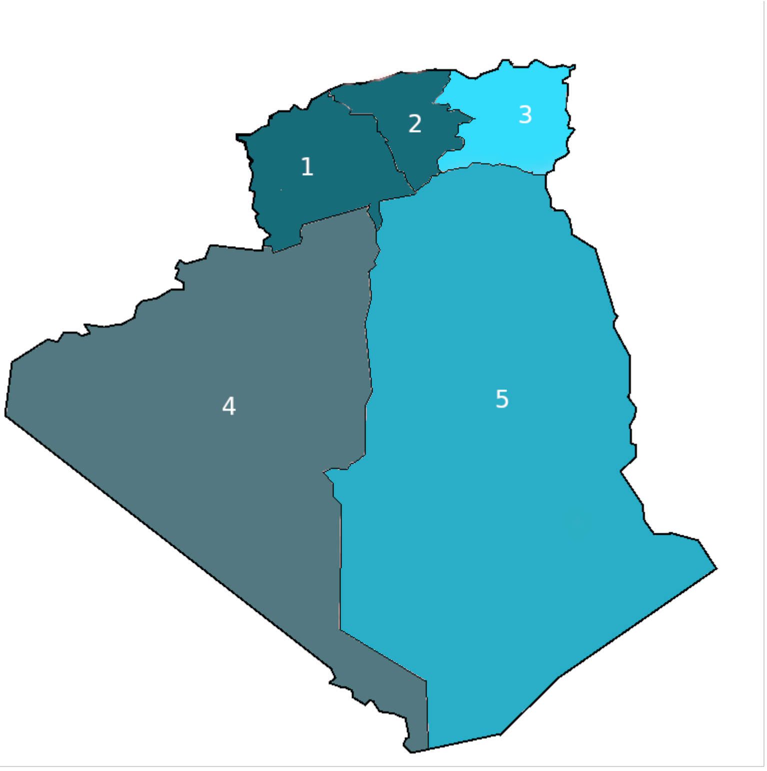


Figure 3.5: Regions

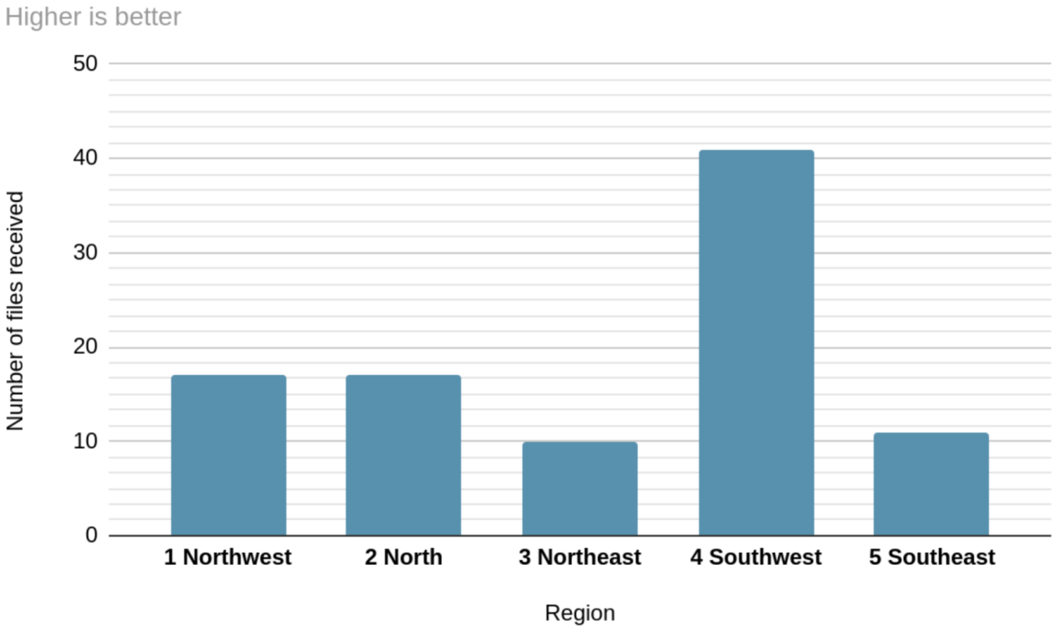


Figure 3.6: Number of files received per region

Southwest comes the first by 41 file, and the last one is Northeast. in addition to 22 files from unknown provinces personal data purpose.

### Part 2 (Twitter)

Each tweet contains 36 properties.We continued to sift data by language. Furthermore, only 160 tweets out of 2000 were able to be translated. As a result(2 hours of working time). We conclude that utilizing tweets to translate is more difficult than using other approaches.

### Part 3 (Youtube)

Using the same idea as Twitter, we focused on Algerian Youtubers[5]. by targeting Algerian commeents and creating the MSA side. we got 560 parallel sentences.

## 3.4 AMSAC discussion

In this section we will present a brief review about the MSA and the Arabic dialects. We will first summarize the brief rules related to the Algerian dialect, with examples. We will also discuss how the Algerian dialect is a variety of languages due to the fact that there are more than one dialect in Algeria, as an adding to the fact that its language is usually spoken and has no standard orthography.

### 3.4.1 Algerian Dialect Study

Algerian dialect has a rich vocabulary that present unique challenge for NLP basics, since this language is usually spoken, it has no established rules to write it, for example a single word or meaning, could have many orthography forms, and the fact of being this dialect different from the MSA it's also different from other Arabic dialects and the other Algerian dialects according to the regions. For example the dialect of the north is different from the south, and the east dialect also differs from the one in the west and center. Other languages, like as French and Berber, have a strong effect on this dialect, the reason that makes it difficult also in the Arab world. In the table 3.1 we summarized a brief characteristics and rules related to this dialect.

[5] https://www.youtube.com/

| / | Rule | Example | English |
|---|---|---|---|
| Sentence structure | VSO | راح سامي لليكون | Sami went to school |
| | VOS | راح لليكون سامي | |
| | SOV | سامي راح لليكون | |
| | SVO | سامي لليكون راح | |
| | OVS | لليكون راح سامي | |
| | OSV | لليكون سامي راح | |
| Negation | ما + verb + ش | ما كتبتش | I didn’t write |
| | | ما لعبتش | I didn’t play |
| | | ما قريتش | I didn’t do my home work |
| Imperative | أ + verb | أشربْ | Drink |
| | أ + verb + ي | أشربي | |
| | أ + verb+ و | أشربو | |
| Present | تـ + verb | تاكل | She eats |
| | يـ + verb | ياكل | He eats |
| | نـ + verb | ناكل | I eat |
| | يـ + verb+ و | ياكلو | They eat |
| | تـ + verb+ و | تاكلو | You eat |
| | نـ + verb+ و | ناكلو | We eat |
| | تـ + verb+ ي | تاكلي | You eat |
| Future | غدوا + present rule | غدوا نخرجو | Tomorrow, we will go out |
| | أومبعد + present rule | أومبعد ناكلو | We will eat later |

Table 3.1: Characteristics of Algerian Dialect

### 3.4.2 Word variation

Algerian dialect is a variety language on it’s own. It has a vocabulary that can have several different forms of a one single word within the different dialects from the north to the east, west, center and south. As the table 3.2 below shows , sometimes one word can have different forms within the different dialects, but other times, the whole word would change.

| MSA | East | West | North | South | Center |
|---|---|---|---|---|---|
| ذهبت | روّحت | رحت | رحت | مشيت | قضيت |
| إدفع لي المال | سلكني | خلصني | خلصني | خلصني | خلصني |
| الآن | توا | درووك | دوكا | دركا | دك |
| اسرع | سرع | غاول | بالخف | ليهليه | فيسع |

Table 3.2: Example of word variations

## 3.5 Statistics

In this section, we are going to show some statistics about AMSAC corpora before normalization and our approach concerning the normalization itself.

### 3.5.1 Before Normalization

| / | Number of letters | Number of words | Vocabulary | AVG word length | AVG sentence length |
|---|---|---|---|---|---|
| AD | 383180 | 70550 | 20214 | 5.43 | 4.81 |
| MSA | 341358 | 68044 | 18518 | 5.03 | 4.64 |

Table 3.3: Statistics before normalization

### 3.5.2 Normalization

The table below shows you the changes applied to AMSAC in order to reduce vocabulary size

| N | Change | Example | | Applied Side | |
|---|---|---|---|---|---|
| | | Before | After | MSA | DA |
| 1 | Remove arabic numbers | ٠ | | yes | yes |
| 2 | | ١ | | | |
| 3 | | ٢ | | | |
| 4 | | ٣ | | | |
| 5 | | ٤ | | | |
| 6 | | ٥ | | | |
| 7 | | ٦ | | | |
| 8 | | ٧ | | | |
| 9 | | ٨ | | | |
| 10 | | ٩ | | | |
| 11 | Remove numeric numbers | 0 | | | |
| 12 | | 1 | | | |
| 13 | | 2 | | | |
| 14 | | 3 | | | |
| 15 | | 4 | | | |
| 16 | | 5 | | | |
| 17 | | 6 | | | |
| 18 | | 7 | | | |
| 19 | | 8 | | | |
| 20 | | 9 | | | |
| 21 | Remove diactilization | أُطْلُبْ العِلْمَ مِنَ المَهْدِ إِلَى اللَّحْدِ. | اطلب العلم من المهد إلى اللحد. | | |
| 22 | Letter unification | ة | ه | No | |
| 23 | | ذ | د | | |
| 24 | | ... | . | | |
| 25 | | أ | ا | | |
| 26 | | إ | ا | | |
| 27 | | آ | ا | | |
| 28 | Sentent structor modification | ؟ - | ؟ - | | |
| 29 | | ؟ - | ! - | yes | |
| 30 | | .- | . - | | |
| 31 | words ending | ى | ا | | |
| 32 | Words starting with | فال | ف ال | No | |
| 33 | | بال | ب ال | | |

Table 3.4: Changes applied to AMSAC in order to reduce vocabulary size

### 3.6 Methodology

Traditionally, Recurrent Neural Networks or (RNN) have been used in Natural Language Processing (machine learning). When a sequence is processed, the hidden state (or ”memory”) that is used to generate a prediction for a token is also sent on, so that it can be utilized to generate the next prediction.

While recurring networks were able to enhance the state-of-the-art in natural language processing, they also had a range of disadvantages:

- RNNs were extremely susceptible to the problem of disappearing gradients. The gradient chain utilized for optimization can be so extensive, especially with extended sequences, that actual gradients are very tiny at the early stages. In other words, the most upstream layers learn essentially nothing, just as with any network impacted by vanishing gradients.
- The same is true of memory. The hidden state is transferred to the next stage of prediction, which means that the majority of available contextual information relates to the short-term view of the model. Therefore, with standard RNNs, models have a long-term memory problem, which is good in the short run but highly poor for the long term.
- Sequential processing occurs. That is, the repetitive network must pass every word in a sentence, which returns a prediction. Because recurring networks can be computationally costly, it can take some time before a prediction is formed on the output. The recurring networks have an inherent difficulty.

Fortunately, many of the above concerns were solved in the 2010s by researching and applying Long-Short Memory networks (LSTMs) and Gated Recurrent Units (GRUs). LSTMs are very resilient to the problem of losing gradients by means of the cell-like structure that holds the memory. Moreover, as memory is now kept independent from the preceding cell output, they are both able to store longer-term memory.

Especially when the attention mechanism was established, where a weighted context vector that weighs the results of all the previous prediction stages is provided in place of the hidden state, long-term memory problems quickly decreased. The only remaining challenge is sequential processing which imposes a considerable resource gap on training a natural language processing model.

#### 3.6.1 Transformers Architecture

Vaswani et al. asserted in an important piece from 2017 that attention is everything you need[25] — in other words, recurring building building blocks are not needed to be truly effective in the NLP tasks of a deep learning model. They presented a novel design, the Transformer, capable of preserving attention during concurrent sequences: all words collectively, not word by word.

The third issue of the above three is that sequences have to be handled sequentially, resulting in a great many computer costs. Parallelism has become tangible with transformers.

Architectures based on transformers are available in many flavors. Researchers and engineers have considerably experimented and brought change based on classic Transformer architecture. But the original architecture of Transformer looks like this 3.8.

As we see the transformers is seperated into two segments.

- An encoder segment that uses source language inputs, creates an incoming embedding for them, encodes positions, calculates where each word needs to work within a multi-context environment and then outputs a put it in perspective.
- A segment of the decoder, using the input of the target language, creates a position embedding for the target language, calculates where the word must be concerned and then combines the output of the decoder with what is produced so far. As a result, a Softmax and hence the argmax class prediction predicts the following token (where each token, or word, is a class).

Consequently, the original Transformer is a classical model sequence to sequence.

### 3.6.2 The translators analogy

Assume our goal is to create a language model that can translate AD text into MSA. In the classic case, using more traditional methodologies, we would learn a model capable of performing the translation directly. To put it another way, we are training one translator to translate AD into MSA. In other words, the translator must be fluent in both languages and comprehend the links between terms in both languages. This will work, but it is not scalable.

Transformers work differently since they are using the architecture of the encoder decoder. Think of it as if two translators work with you. The first translator can translate AD into a global intermediate language. Another translator is able to translate it into MSA. You will nonetheless allow translations pass through the intermediate language first for every translation task. But it's scalable as well: for instance, we may utilize the intermediate language to train a text resumed model.

#### The encoder segment

A Transformer's encoder section is in charge of transforming inputs into some intermediary, high-dimensional representation. Visually.

- *Input Embedding* are scripts that translate tokenized inputs into vector format so they can be used.
- *Positional Encoding* alter the vector outputs of the embedding layer somewhat, providing positional information to these vectors.
- *The actual encoder segment*,that learns to display the input vectors attended.
  - *Multi−head attention segment*, It does multi-head self-attention, adds the residual connection, and then normalizes the layers.
  - *Feed Forward Segment*,For each token, the encoder output is generated.
  - The encoder section can be repeated a Nth number of times.

### N times encoder segment

- *Multi−Head attention block*: There's a multi-head attention block moving on. This block enables self-attention during each sequence.
- *Feed Forword block*: It generates a dmodel-dimensional and hence 512-dimensional vector that encodes the token after generating attention for each token (word).
- *Residual connections*: A connection that does not flow through a complex block is referred to as a residual connection. Two residual connections can be seen here: one from the input to the first Add  Norm block, and another from there to the second block. Because gradients can flow freely from the end of the model to the beginning, residual connections allow the models to optimize more efficiently.
- *Add & Norm blocks*: The output from either the Multi-head attention block or the Feed-forward block is combined with the residual in these blocks, and the result is then layer normalized.

### The decoder segment

This Transformer component is in charge of turning the intermediary, high-dimensional representation into predictions for output tokens.

- *Output Embeddings*: which, like the embeddings used for the inputs, converts tokenized outputs into vector format. The only change here is that the outputs are one position to the right.
- *Positional Encodings*: which, like the input positional encodings, modifies the vector outputs of the embedding layer somewhat, adding positional information to these vectors
- *Actual decoder segment*: Composed of two things.
  - *Masked multi−head attention segment* which executes multi-head self-attention on the outputs in a disguised manner, so that positions are solely dependent on the past.
  - *multi−head attention segment* The model learns to associate encoded inputs with desired outputs by performing multi-head self-attention on a mixture of (encoded) inputs and outputs.
- The feed forward segment, which processes each token individually.
- Finally, a linear layer is used to generate logits, and a Softmax layer is used to generate pseudoprobabilities. We know which token to take and add to the tokens already predicted by obtaining the argmax value of this prediction.

### N times Decoder Segment

The functioning of the first two components of the decoder segment was identical to that of the first two components of the encoder segment.

- *A masked multi−head attention segment* where the model learns which prior tokens it needs attend given some token by applying self-attention to (masked) outputs.

- $A\ multi{-}head\ attention\ segment$ where self-attention is applied to encoded inputs (acting as queries and keys) and the combination of masked multi-head attention outputs / input residual is the gateway via which encoded inputs and target outputs are mixed.
- $A\ feedforward\ segment$ which is applied to each token as it is handed along.

Finally, a linear layer and a Softmax activation function are included as an appendix. These will take the decoder segment output and convert it into a logits output and a pseudo probability output that assigns probabilities to each of the potential token outputs given the logit values. We can discover the most likely forecast here by just taking the argmax value from these results.

$$Attention(Q, K, V) = \frac{Softmax(QK^T)}{\sqrt{d^K}}V$$

Figure 3.7: Attention

- $Q$ is a matrix that contains the query (vector representation of one word in the sequence).
- $V$ are the values, which are again the vector representations of all the words in the sequence.
- $K$ are all the keys (vector representations of all the words in the sequence).

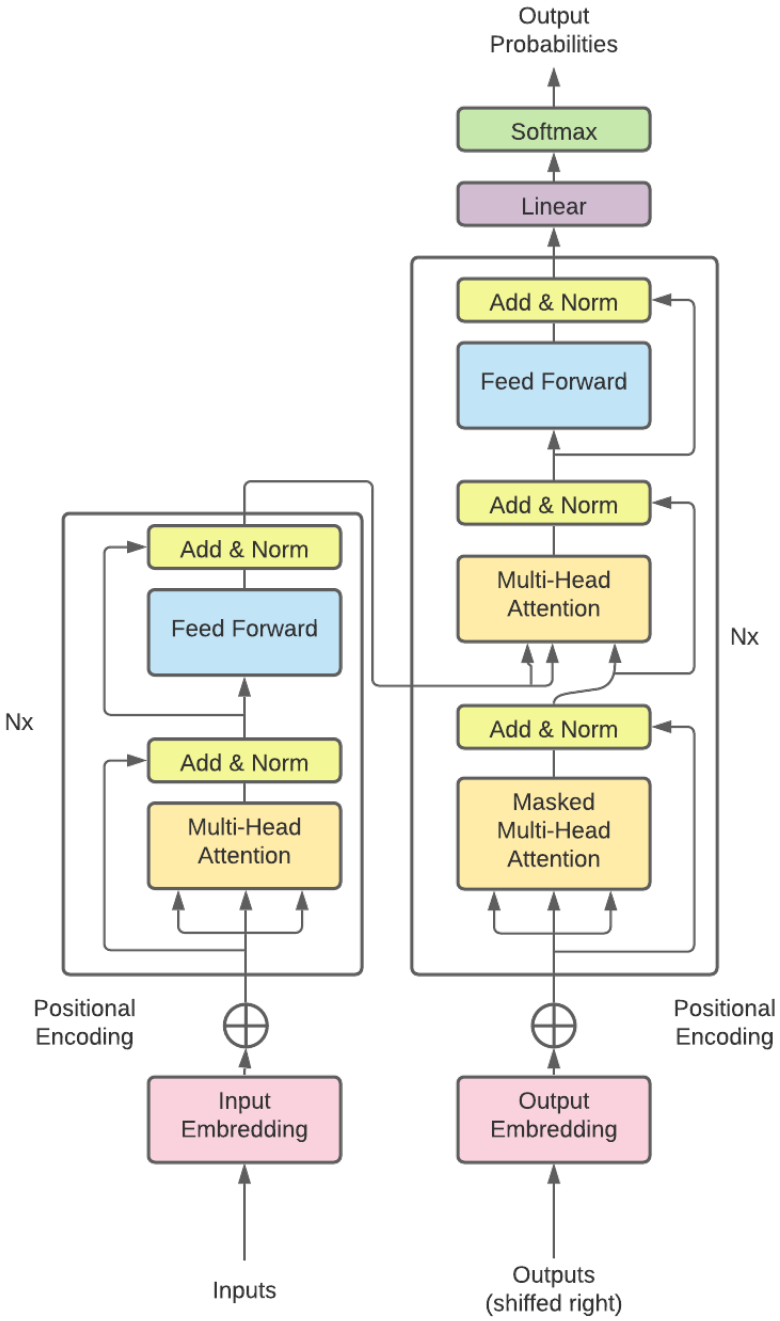


Figure 3.8: Transformers Architecture

# Chapter 4

# Implementation and Results

## 4.1 Introduction

In this chapter we will present implementation and results of this project, including the development environment and the evaluation method for our model. We will also present our AMSAC corpus, the experiments and the final tests. Finally we will present the different GUI of our proposed model LAHDJA and the conclusion.

## 4.2 Used tools

In this section, we represent different tools used for our experiments.

| N | System \| Tool | Description | Version |
|---|---|---|---|
| 1 | Ubuntu[1] | - Linux Operating System | 18.04LTS |
| 2 | VS Code[2] | - IDE | 1.56.2 |
| 3 | BPE[3] | - Sub Word NMT | 0.3.7 |
| | | - Segment text into subword units | |
| 4 | Buckwalter[4] | - Transliterate arabic text to alphabetic letters | 0.6.0 |
| 5 | Fairseq[20] | - Sequence modeling toolkit using Pytorch | 0.10.2 |
| | | - customizable toolkit for creation Models of | |
| | | - Translation | |
| 6 | Python[5] | - Object Oriented Programming language | 3.8.x |
| 7 | Dart[6] | - Object Oriented Programming Language | 2.13 |
| 8 | Flutter[7] | - Google's UI toolkit for web,desktop and mobile apps | 2.2 |
| | | - Based on Dart | |

Table 4.1: System and Tools

[1] https://ubuntu.com/

[2] https://code.visualstudio.com/

[3] https://pypi.org/project/subword-nmt/

[4] https://pypi.org/project/lang-trans/

[5] https://www.python.org/

[6] https://dart.dev/

[7] https://flutter.dev/

**BLEU(Bilingual evaluation understudy) method for LAHDJA's evaluation** BLEU method is an automatic tool for machine translation's evaluation, proposed on July 2002. It proved her benefit by deriving her judgment from human understanding. This approach based on the rule:[21]

$$BLEUScore = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where:

$$BP = \begin{cases} 1, & \text{if } c > r \\ exp(1{-}\mathrm{r/c}), & \text{if } c \leq r \end{cases}$$

and:

$$p_n = \frac{\sum_{C \in Candidates} \sum_{n-gram \in C} Count_{clipped}(n-gram)}{\sum_{C' \in Candidates} \sum_{n-gram' \in C'} Count_{clipped}(n-gram')}$$

$N$ : Length of sentence

$w_n$ : Positive weights

$c$ : Length of the candidate translation

$r$ : The effective reference corpus length

## 4.3 AMSAC

After normalizing the data in our corpora we got results shown in the table below. In total we have 14655 parallel sentences. We passed MADAR[4] by 2655 sentences and also PADIC[17] by 8255 sentences. AMSAC now is the biggest parallel corpora which contain a Dialect and MSA translation.

| / | Number of words | (+/-)% | Vocabulary | (+/-)% | AVG word length | (+/-)% | AVG sentence length | (+/-)% |
|---|---|---|---|---|---|---|---|---|
| AD | 62136 | -11.93 | 17222 | -14.80 | 5.01 | -7.73 | 4.69 | -2.49 |
| MSA | 64100 | -5.80 | 17442 | -5.81 | 5.31 | 5.57 | 4.85 | 4.53 |

Table 4.2: Statistics after normalization

## 4.4 Experiments

In this section we will present our experiments on AMSAC.Firstly, we present used tools in our model "LAHDJA".

### 4.4.1 Impact of BPE and vocabulary on BLUE Score results

This experiments were applied on part of AMSAC. the goal behind this experiment was getting the impact of BPE (vocabulary size) and layers of the module on BLEU score result. The number of sentences used in this part was 13.000 sentences. 10.000 for training, 1.000 for validation, 2.000 for testing. on this expirement the normalization were appled for both sides(AD and MSA)

### Used parameters

This Table shows the parameters that we modified in Fairseq tool in order to built our model.

| Parameters | Description |
|---|---|
| Learning Rate | Determines how quickly the model will respond to newly measured errors. |
| Encoding Layers | How much layer for encoding process |
| Decoding Layers | How much layer for decoding process |
| Head Attentions | How much time the sentence pass in the function to be evaluated |
| Epoch | Number of epochs |
| Max Tokens | Token per part |
| Batch Size | How much sentences in a single step |
| Drop Out | Speed of drop out |
| Layer Drop | Layer drop |

Table 4.3: Used parameters in Fairseq

### Giving parameters values

- Vocabulary size: changes from 300 up to 3500.
- Layers: from 2by2 until 4by4.
- Learning rate: 4.
- Head-attention: 4.
- Batch size: 512.
- Epochs: 50.
- Tokens: 400.
- Layer drop: 0.004.
- Drop out: 0.015.

### Impact of number of layers on Blue score results (with bpe ”BLEU” and without bpe ”BnT”)

| / | 2x2 | | | 3x3 | | | 4x4 | | |
|---|---|---|---|---|---|---|---|---|---|
| VOCAB | BLEU | PPL | BnT | BLEU | PPL | BnT | BLEU | PPL | BnT |
| 200 | 19.55 | 12.88 | 10.38 | 19.32 | 12.91 | 9.51 | 19.44 | 13.06 | 10.25 |
| 300 | 18.14 | 15.19 | 10.8 | 17.81 | 15.55 | 9.33 | 17.24 | 15.85 | 10.38 |
| 400 | 16.71 | 17.31 | 9.84 | 16.38 | 18.28 | 10.1 | 15.92 | 18.28 | 10.14 |
| 500 | 15.41 | 20.23 | 10.4 | 15.12 | 20.47 | 10.49 | 12.76 | 21.01 | 10.35 |
| 1500 | 10.87 | 44.75 | 10.6 | 11.14 | 46.76 | 11.15 | 10.61 | 49.49 | 10.69 |
| 2000 | 9.27 | 69.76 | 10.47 | 8.95 | 74.61 | 9.51 | 7.55 | 81.48 | 9.73 |
| 3000 | 8.88 | 83.02 | 9.88 | 8.88 | 87.96 | 10.11 | 9.13 | 93.03 | 8.95 |
| 3500 | 7.92 | 94.42 | 9.11 | 8.81 | 100.07 | 10.26 | 7.26 | 114.71 | 9.91 |

Table 4.4: Impact of number of layers on Blue score results (with bpe ”BLEU” and without bpe ”BnT”)

**Discussion**

- BLEU results will increase if we have a smaller vocabulary size.
- The impact of BPE in this experiments is limited.
- Perplexity increases if the vocabulary size increased.

**Reasons**

- BnT (BLEU score without BPE tokenizer) results are close because of how BLEU is calculated
- The characteristics of the corpora and variation of sentences increases the difficulty of learning the data
- Lack of consistency in the sentences

### 4.5 Final Tests

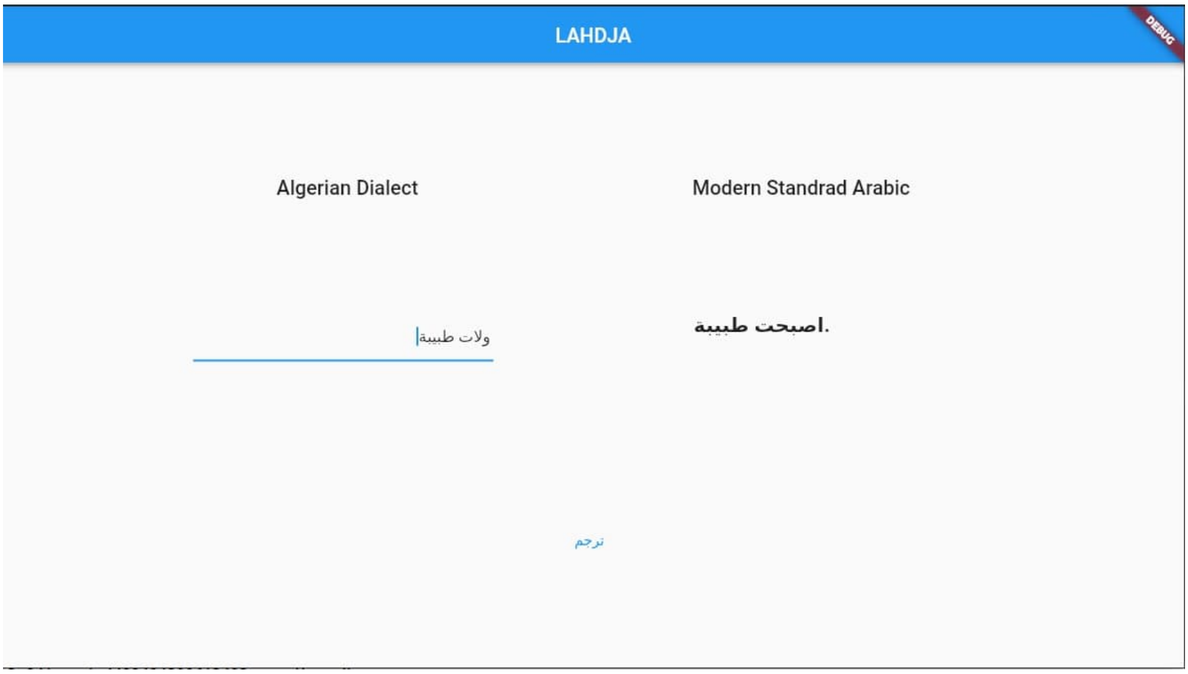
After completing the version 1.0 of AMSAC. We tried multi ways to improve our results by limiting the length of sentences, using BPE and using Buckwalter transliteration. We got the below results

As we see in this results, the impact of BPE and Buckwalter is remarkable in which the results has been changed from 7.74(the worst) to 15.13 (the best results). Our model has achieved 15.13 BLEU score which is more than the Meftouh[17] results by 0.03 in Algerian Dialect Modern Standard Arabic translation.

## 4.6 Our Application

In this section we will present the LAHDJA interface we built. It requires the operating system Linux and the availability of the Flutter library to be run.

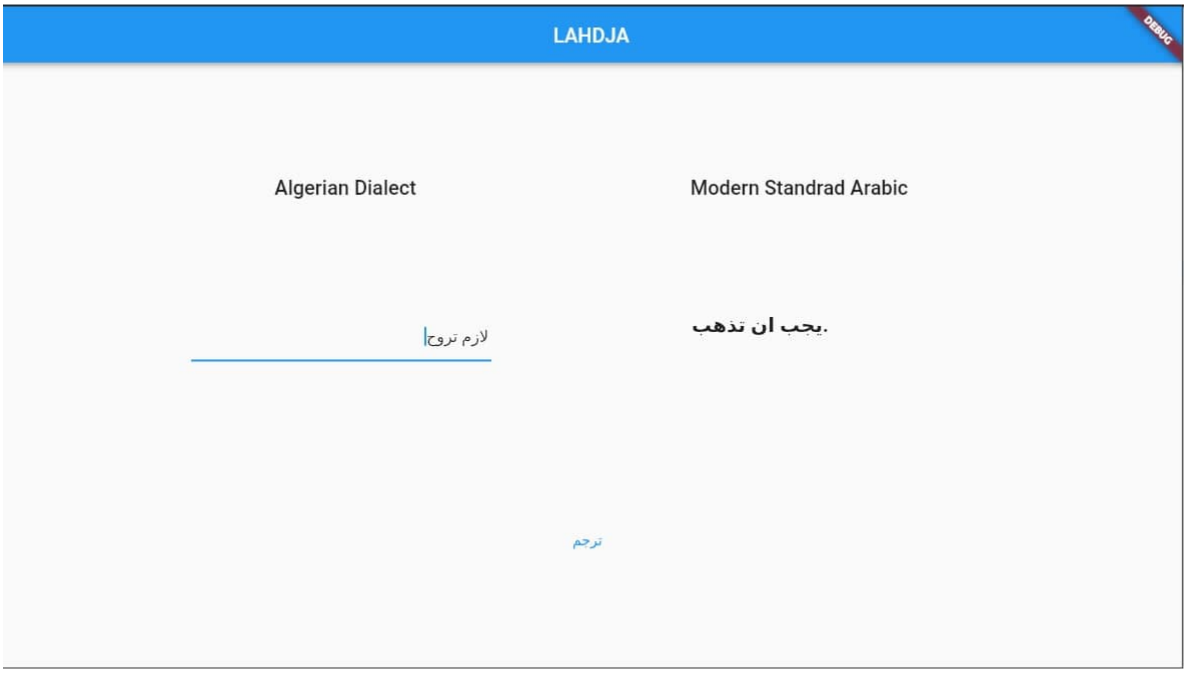
LAHDJA application contain only one interface with a simple input/output components. The uses the Faireseq-interactive tool directly to get the translation.



ولات طبيبة

اصبحت طبيبة.

Figure 4.1: LAHDJA GUI



لازم تروح

يجب ان تذهب.

Figure 4.2: LAHDJA GUI

## 4.7 Conclusion

With regards to our AMSAC corpora of over 14k parallel sentences(the largest Algerian dialect corpora until now) which we built, we achieved the best results in translation of Algerian dialect until now, compared to the Meftouh[17] model with a 15.13 BLUE score. In this chapter, we presented implementation and results of our approach to build the AMSAC corpus and LAHDJA model, including statistics, experiments and final tests. We have also presented the development environment and the LAHDJA application.

# Conclusion

Arabic is the official language of the Arab world used in official domains such as news talks, newspapers and for learning in school. It is the standard form language overall in Arabic countries. In contrast, the daily spoken language used are the dialects.

Arabic dialects are non-standard variants of Arabic that are usually spoken and increasingly written throughout the Arab world. These dialects have a lack of uniform orthographies and have been classified as under-resourced languages, the thing that presents a barrier for natural language processing applications. While natural language processing (NLP) researchers and developers are more likely to focus on the standard form of any language, more research started to be focused on to meet the demands of non-standard variations and dialects to translate these dialects to the standard form. This area of studies is a multipurpose research, it can help to control and classification in the field of sociology and context categorization as well as in the commercial and communication domains and make it easy to the non-native or non Arab people to deal with one single language which is the standard form of the Arabic instead of multi dialects.

In this project, we first presented our AMSAC corpus (Algerian dialect Modern Standard Arabic Corpora) , a parallel corpus for Algerian dialect with the MSA. AMSAC corpus is a collection of 14655 sentences containing a vocabulary of 17222 of Algerian dialect words and 17442 of MSA words in parallel. We have also presented the steps of collection and normalization and its statistics to achieve the good results at. We have also presented our proposed model LAHDJA, a translator model for the Algerian dialect to the modern standard Arabic. The LAHDJA model has achieved better results than the Meftouhe in Algerian Dialect Modern Standard Arabic translation. We have presented the conceptions of the translation model LAHDJA, the implementation and methodology and finally the final tests and their results.

Our immediate next steps are to make the application available for mobile phones. We will also work on extending the size of the AMSAC corpus by adding more sentences for achieving the best results always.

# Bibliography

[1] E. Abuelyaman, L. Rahmatallah, W. Mukhtar, and M. Elagabani. Machine translation of arabic language: challenges and keys. In *2014 5th International Conference on Intelligent Systems, Modelling and Simulation*, pages 111–116. IEEE, 2014.

[2] L. H. Baniata, S. Park, and S.-B. Park. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational intelligence and neuroscience*, 2018, 2018.

[3] A. Baumgartner-Bovier. La traduction automatique, quel avenir? un exemple basé sur les mots composés. *Cahiers de linguistique française*, 25:274, 2003.

[4] H. Bouamor, N. Habash, M. Salameh, W. Zaghouani, O. Rambow, D. Abdulrahim, O. Obeid, S. Khalifa, F. Eryani, A. Erdmann, et al. The madar arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[5] P. Bouillon. *Traitement automatique des langues naturelles.* De Boeck Supérieur, 1998.

[6] J. Chandioux et al. Histoire de la traduction automatique au canada. *journal des traducteurs*, 22(1):54–56, 1977.

[7] M. A. Chéragui. Theoretical overview of machine translation. In *ICWIT*, pages 160–169. Citeseer, 2012.

[8] D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012.

[9] M. Cori and J. Léon. La constitution du tal. *Traitement Automatique des Langues*, 43(3): 21–55, 2002.

[10] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2011.

[11] Y. Fukumochi. A way of using a small mt system in industry. *the 5th Machine Translation Summit*, 1995.

[12] S. Harrat, K. Meftouh, M. Abbas, W.-K. Hidouci, and K. Smaili. An algerian dialect: Study and resources. *International journal of advanced computer science and applications (IJACSA)*, 7(3):384–396, 2016.

[13] W. J. Hutchins and H. L. Somers. *An introduction to machine translation*, volume 362. Academic Press London, 1992.

[14] H. Kaji. Hicats/je: a japanese-to-english machine translation system based on semantics. In *Machine Translation Summit*. Citeseer, 1987.

[15] Y. Lepage and E. Denoual. Aleph: an ebmt system based on the preservation of proportional analogies between sentences across languages. In *International Workshop on Spoken Language Translation (IWSLT) 2005*, 2005.

[16] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, and K. Smaili. Machine translation experiments on padic: A parallel arabic dialect corpus. In *The 29th Pacific Asia conference on language, information and computation*, 2015.

[17] K. Meftouh, S. Harrat, and K. Smaïli. Padic: extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*, 2018.

[18] P. P. Monty. Traduction statistique par recherche locale. 2010.

[19] H. Mubarak. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), OSACT2018 Workshop*, pages 49–53, 2018.

[20] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[22] F. Sadat, F. Mallek, M. M. Boudabous, R. Sellami, and A. Farzindar. Collaboratively constructed linguistic resources for language variants and their exploitation in nlp application–the case of tunisian arabic and the social media. In *Proceedings of workshop on Lexical and grammatical resources for language processing*, pages 102–110, 2014.

[23] M. A. Sghaier and M. Zrigui. Rule-based machine translation from tunisian dialect to modern standard arabic. *Procedia Computer Science*, 176:310–319, 2020.

[24] K. Shaalan, S. Siddiqui, M. Alkhatib, and A. A. Monem. Challenges in arabic natural language processing. *Computational Linguistics*, 2019.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[26] F. Yvon. Une petite introduction au traitement automatique du langage naturel, support de cours. *Ecole Nationale Supérieur des télécommunications*, 2007.

[27] R. Zbib, E. Malchiodi, J. Devlin, D. Stallard, S. Matsoukas, R. Schwartz, J. Makhoul, O. Zaidan, and C. Callison-Burch. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59, 2012.

[28] J. Zhang, C. Zong, et al. Deep neural networks in machine translation: An overview. *IEEE Intell. Syst.*, 30(5):16–25, 2015.