ASJP Algerian Scientific Journal Platform

Revue El Omda En linguistique et analyse du discours

Issn:2572-0058/E-issn:1969-2676

https://www.asjp.cerist.dz/en/PresentationRevue/485

Volume: 06/ N°:01 (2021) p 477- 485

Computational Arabic linguistics: An example of Natural Language Processing

Abdelkader BENSAFA*

Abdelkaderamine08@gmail.com

Abstract ;

With the rapid change of today's world, many disciplines witnessed lot of –sometimes radical changes; and so is the study of human language. One of the agents contributing in the process is Information and Communication Technologies in general and computers in particular. This last helped in the development of NLP applications. Natural language processing (NLP) is viewed as the ability of a computer program to understand human speech as it is spoken. All of which resulted in what is called now computational linguistics. The aim of this paper is to explore in the notion of Computational Arabic linguistics and prisizely Natural Lnaguge Processing (NLP) by adressing the following questions: What was done? And what is missing? All ofwhich is to pave the way for our university students to help them in the process of decoding the current practices and the future directions regarding the Arabic language, besides facilitating cultural communication with other.

Article info

Received
14/07/2021
Accepted
09/11/2021

Keyword:

- ✓ *ICT*
- ✓ *Computer*
- ✓ *Language Studies*
- ✓ *NLP*
- ✓ *Computational Linguistics*

*Corresponding author

Introduction

No one can neglect the crucial changes brought by the integration of information and communication technologies (ICT's) to our daily life. The process of development has been accelerated and so is the domain of education. Amongst the spheres of education, linguistics makes profit from that integration of ICT to bridge the gap between the old fashion of studying or viewing language and the 21st demands and trends. As we all know, linguistics is the scientific study of language based of course on the rules of science (observing, problem forming, hypothesizing, testing, and conclusion or problem solving).

To render it more scientific, linguists shifted their attention to the use if ICT in studying language. To this end, concepts such as Natural Language Processing (NLP), Computational Linguistics (CL) Artificial Intelligent (AI) will be highlighted in little more details throughout this paper. It should be stressed here that the current research is a review of literature based on a descriptive approach which is part of the deductive paradigm. In other words, it will serve as a point of departure for future empirical or inductive research.

1. **Information and communication technology (ICT)**

The existing literature about ICT is very rich as many researchers and educationalists give it hight emphasize. In here, a bird eye view will be provided. Believing that research is about defining the unknown and redefining the known, I am going to define ICT focusing on the following points:

1. **Information**

To understand the term information, it is worth mentioning to have a look at the following pyramid

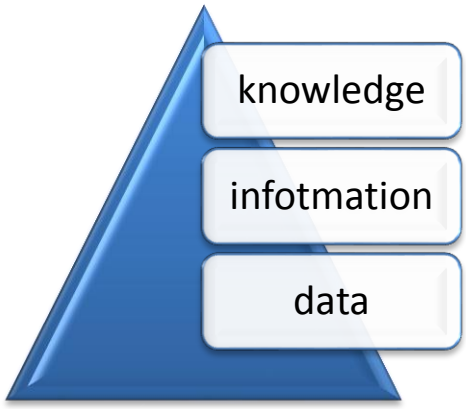

Figure 1.1 the Down-Top Process of Inquiry
(Adopted from Rowley 2007)

Based on the figure mentioned above, data is regarded as the primary source or the ground of any inquiry. It is characterized to be abstract, general, and with no context. Let us consider the following numbers 17042020. The first reading of those numbers, nothing can be grasped or understood.

Now, if data is contextualized, here it becomes information as the human mind will decode what was encoded. So, again let reconsider the previous numbers as

follows 17/04/2020, automatically one can understand that it is a date of birth, a phone number, or a bank account...etc. In sum, once data is contextualized it becomes information. To this end, if one's horizon is widened, the pieces of information will enlarge one's knowledge which is the highest status of the pyramid mentioned previously.

In brief, data is the primary source or the ground, once it is contextualized it will result in informing us about something (information) and in return, information will result in knowledge (an increased understanding of the living world).

1. **Communication**

Communication is the process of exchanging ideas, beliefs, thoughts, and knowledge. This process falls under 2 categories, 5 types, and 3 models.

1. **Categories**

The process of communication is- with no doubt- one of the following categories**:** Verbal using spoken or written language; or Non-Verbal using signs, gestures, and facial expressions, i.e., Body language.

2. **Types**

When speaking about the different types of communication, Dance and Larson 1972) listed the followings:

a. **Extra-personal communication:** a communication between a human being and a non-human being (animals). For example, a man with his dog, cat, or bird.

b. **Intra-personal communication:** a communication between a human being and himself. It is likely to appear or happen to all of us as humans where we talk explicitly or implicitly with ourselves.

c. **Inter- personal communication:** a communication between human beings despite their level of education, race, and ethnicity. In here, concepts such as formal, non-formal, and informal norms are of value.

d. **Organizational communication:** inter- personal communication will certainly have a context or an organization. For example, a teacher with his students in a classroom, a boss with his workers in a factory or institution...etc. what should be buried in mind regarding this type is the notion of organization**.**

e. **Mass communication:** a communication between human beings across distances or beyond the boundaries of a village or country. This type makes use of technology i.e., internet, social media, TV's, Radio...etc.

1.1.1 Models

Linguists and applied linguists join both categories and types of communication under the following models. The simplest model (a)

Figure 1.2 the Linear Model of Communication
(Adapted from McCornack and Ortiz 2017)

However, that model was criticized to miss important details concerning the process of communication. This is why; the next model (b) was seen to have clear

insights.

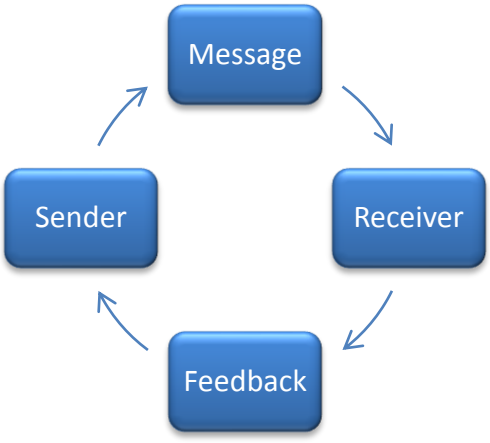

Figure 1.3 the Cycle Model of Communication
(Adapted From U.S. Government Printing Office 1995)

In this model the notion of Feedback was added for the simple reason that the process of communication is not only a one-way process but rather a two-way one. This means that the message encoded by the sender will be decoded by the receiver; in return, the receiver will also encode his feedback to be decoded by the sender i.e., the positive or the negative reaction and /or interaction. The last model **(c)** takes the **(b)** model as a framework and adds the concept of Noise.

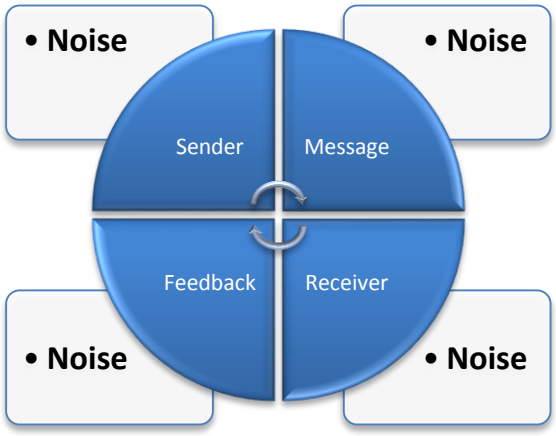

Figure 1.4 the Noise Model of Communication
(Adapted From Littlejohn and Foss 2008)

1.2 Technology

The last part of ICT is technology. Technology refers to the use of the different devices science has come up with. It facilitates life and is an ongoing or endless process. For example, phones, cars, TV's and mainly computers. Computers paved the way to Natural Language Processing (NLP). This last will be highlighted in the following part.

2. Natural Language Processing (NLP)

To the best of my knowledge, NLP can be given the status of being a discipline. This conclusion is due to the fact that the four (4) criteria researchers agreed on to have a discipline are present:

a. Researchers or experts embarked in the field of NLP
b. Publications including books, journals, and papers
c. Students whose desire is to be competent in the field
d. Organizations including governmental or non-governmental institutions whose ultimate aim is to bring both experts and students under the same organization to facilitate their path.

2.1 NLP defined

Natural Language Processing also known as human language Technology (HLT) or Natural Language engineering (NLE) is part of Computational Linguistics (will be discussed later on) where the state of art is to process information contained in natural language texts.

In the same vein, NLP is about the use of machines to analysing, understanding, and generating human language just like humans do. This task is realized by applying computational techniques to language domain. But do really machines understand human language? The answer to this question turns around our exact definition of the word "understand" (Garbade 2018)

Understanding is the ultimate goal. However, one does not need to fully understand to be useful. In other words, the role of machines in general and computers in particular in NLP is to make computers learn our language rather than or as we learn theirs. This leads to say that NLP starts off as a branch of artificial Intelligent to explain the logistics theories to build systems that can be of social use or value. (ibid)

In the same line of thought, one can come up with the conclusion that NLP is a multidisciplinary as it borrows from other disciplines mainly linguistics, psycholinguistics, cognitive science, computing and statistics.

2.2 Applications of NLP

Among the applications of NLP one can list the followings:

a. Machine Translation and Database Access
b. Information retrieval or selecting from a set of documents the one that are relevant to a query
c. Text organization i.e., sorting texts into fixed topic categorization
d. Extracting data from a text i.e., converting unstructured texts into structured data
e. Spoken language control systems
f. Spelling and grammar checkers

As mentioned previously, NLP is better seen within the so-called Computational linguistics. The following section will be devoted to Computational Linguistics in general and Computational Arabic Linguistics in particular.

3. Computational Linguistics (CL)

As it was preceded with NLP, definition and applications of CL will be given.

3.1 Definition

Computational linguistics or CL is a discipline between linguistics and computer sciences which is concerned with the computational aspects of human language faculty. In other words, it belongs to the cognitive sciences and overlaps with the field of Artificial Intelligence (AI). (Uszkoreit, 2000)

Hand in hand with the definition of CL, both applied and theoretical components do exist in Computational linguistics as it is a branch of Computer science aiming at computational models of human cognition.

3.2 Applications of Computational linguistics

This part summarizes -but not limited to- the main applications found in CL.

Table 3.1 Applications of CL

Application	**Application**
Text normalization/segmentation	Machine translation
Semantic role labelling	Parsing
Application	**Application**
Automatic word pronunciation prediction	Morphological analysis
Dialog systems	Summarization
Topic detection	Computer-aided language learning (CALL)
Language modelling for automatic speech recognition	Bioinformatics
Text retrieval	Transliteration
Word-class prediction: e.g., part of speech tagging	

4. Computational Arabic Linguistics (CAL)

When talking about what was done in the field of CAL, many examples or works can be cited, but first let us have a look at the following definition (Fraghly 2010: np) summarizing the rationale behind CAL:

> Arabic is an existing- yet challenging- language for scholars as many of its linguistic properties have not been fully described. So, Computational Arabic Linguistics or sometimes called Arabic Computational Linguistics documents the recent works of researchers in both academia and industry who have taken up the challenge of solving the real- life problems posed by an understudied language.

So, one can easily depict that when it comes to Arabic, research is still in its beginning and much will be done later on. Here are some examples:

4.1 Arabic Machine Translation Systems

Since it was developed, Arabic machine translation has been subject to description and evaluation. The first English to Arabic machine translation system was developed in the late seventies by Weidner Communications Inc. which was located in Provo, Utah. The system was developed following the Direct Method which aimed to produce fully automated Arabic translations of unlimited English source documents but did not limit the translation to a specific domain. (Izwaini 2006)

As in all other MT systems that adopted the direct method, it was designed for a specific pair of languages: English as the source language and Modern Standard Arabic as the target language. According to Bassiouney and Katz (2012:45)

> The system consisted of two main stages: analysis of the source language and generation of the target language. The analysis of English was oriented to enable the correct generation of target language expressions employing a large bilingual dictionary as well as a dictionary for idiomatic expressions.

Based on the quotation above, the vocabulary and syntax of English was not analysed in depth and only to the extent required to generate Arabic equivalents. Thus, the system was unidirectional and did not perform deep syntactic or semantic analysis of the source language. The system was commercially utilized by Omnitrans of California Inc. which used it for the purpose of translating the Encyclopaedia Britannica into Arabic. This project was not completed for lack of funding. (Bassiouney and Katz 2012)

4.2 Speech Recognition

Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone.

According to Satori et al (2007:2):

> The first works on Arabic ASR has concentrated on developing recognizers for modern standard Arabic (MSA). The most difficult problems in developing highly accurate ASRs for Arabic are the predominance of non-discretised text material, the enormous dialectal variety, and the morphological complexity

Although Arabic is currently one of the most widely spoken language in the world, there has been relatively little speech recognition research on Arabic compared to the other languages.

4.3 Mention Detection

Mention Detection (MD) is a basic task of information extraction. Besides the identification and classification of textual references to objects/abstractions (i.e., mentions). These mentions can be either named (e.g., Mohammed, John), nominal (city, president) or pronominal (e.g., he, she).
For Arabic MD, Arabic adopts a very complex morphology, i.e., each word is composed of zero or more prefixes, one stem and zero or more suffixes. Consequently, the Arabic data is sparser than other languages, such as English, and it is necessary to "segment" the words into several units of analysis in order to achieve a good performance. (Banajiba et al 2010)

Conclusion

It may be deduced that the field of computational linguistics is relatively young compared to other areas of study and particularly more so in Arabic. With the sole purpose of bettering machine language production and interpretation, computational linguistics found its footing with the advent of technology. Understandably, the advances of Arabic computational linguistics are still in the early stages.

References

Bassiouney R & Katz E. G (eds.), 2012 Arabic Language and Linguistics. Georgetown University Press Round Table on Languages and Linguistics series, 246 pages, ISBN: 9781589018853 (1589018850).

Benajiba, Y & Zitouni, I (2010) Arabic Word Segmentation for Better Unit of Analysis
Dance, E. X. and Larson, C, E. (1972), Speech Communication: Concepts and Behaviors New York, NY: Holt
Farghaly A (2010) Arabic computational linguistics. Stanford USA

Garbade M, J (2018) A Simple Introduction to Natural Language Processing

Izwaini, S. (2006) Problems of Arabic Machine Translation: Evaluation of Three Systems. Proceedings of the International Conference at the British Computer Society (BSC), London.

Knapp M. & Dary J. (2000) Handbook of Interpersonal Communication 3rd Ed (p102-129) Newbury Parkl CA: Sage'

Littlejohn, S.W. & Foss, K.A. (2008) Theories of Human Communication, 9th edition. Belmont, CA: Thomson Wadsworth

McCornack, S and Ortiz, J. (2017) Choice & Connections: An Introduction to Communication
Rowley, J (2007). "The wisdom hierarchy: representations of the DIKW hierarchy". Journal of Information and Communication Science. 33 (2): 163–180

Satori, H. , Harti, M & Chenfour, N. (2007). Introduction to Arabic Speech Recognition Using CMUSphinx System. CoRR. abs/0704.2083.

U.S. Government Printing Office (1995)"Global communications : opportunities for trade and aid" U.S. Congress, Office of Technology Assessment. (OTA-ITC-642nd Ed.)

Uszkoreit, H (2000) What Is Computational linguistics