Proceedings of Engineering & Technology (PET)
pp. 784-789
Copyright IPCO-2016

Semantic Error Detection in Arabic Language Using Ontology

Asma Ksiksi   Chekib Gmati   H. Amiri
Signal, Images and Technologies of Information laboratory (LR-SITI)
National Engineering School of Tunis (ENIT)
Tunis, Tunisia
asma.ksiksi@gmail.com, chekibgmt2007@yahoo.fr, hamidlamiri@gmail.com

***Abstract*— In this article, we present a method for automatic detection errors from Arabic text those mistakes become more and more frequent and can break the semantic consistency of the sentence. The necessity of an automatic detection of the committed mistakes becomes more and more obvious. Ontology is widely used in the analysis of Arabic text; it will represent the essential foundation of our approach. In fact, the different formalisms available in the ontology will provide an efficient element to support the complexity of the Arabic language. Proceeding from the fact that a grammatical error is a modified correct rules. We will take advantage of representations of correct rules in the ontology to achieve the detection of semantic errors.**

Keywords*— *Ontology, natural language processing, detection of errors, rules, grammars , arabic text.

I. Introduction

Arabic language is used by millions of people in northern African and in twenty Middle East countries. It is the fifth spoken language in the world [1] [2]. It represents the most contemporary spoken Semitic language today with more than 300 million speakers [3].

This language is recognized by her richness [4], It has become increasingly utilized used in internet. Arabic became then an essential means of communication which made writers and linguists develop basic standardization of Arabic grammar. This development was accompanied by rapid and deep changes especially in syntax and lexical enrichment. This development which extends to the digital world requires automatic processing of Arabic language. linguistics. Among them we mention the autom atic translation [5], the automatic extraction of opinions and emotions [6], and the automatic summarization [7], etc.

On the other hand, the wealth and the linguistic complexity of the Arabic language require automatic analysis to make corrections, to locate errors and several other treatments that affect the grammatical and lexical coherence of Arabic language.

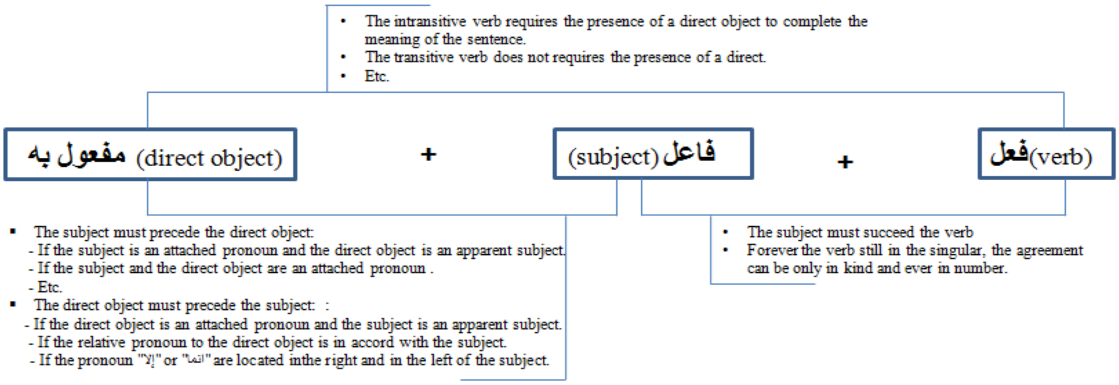
In this paper, we address the problem of errors that can produce words lexically valid but semantically invalid.

II. THE CHARACTERISTICS OF THE ARABIC LANGUAGE

The Arabic language is distinguished from other language by a wealth morphological and syntactic system. In fact, the construction of lexical units and their transformation in accordance with the desired meaning is so varied. However, the grammatical coherence is ensures by analyzing the units' position and case marking.

In what follows, we will present some grammatical rules highlighting the morpho-syntactic relation of the Arabic language. We present in the Fig. 2 some rules that are specific to the verbal phrase.

In the Arabic language, we find properties that are decisive to express the semantic consistency: The definite noun « إسم معرّف» : the name must be preceded by the definite article ال ; The negation: it can be ensured by negation article such as ما, لا, etc; The coordinating conjunctions «العطف »: connects two words, phrases or clauses together; The adjective «النَّعْت»: come after the nouns. It must agree with the nouns in terms of number, gender, state of definiteness and grammatical case.



- The intransitive verb requires the presence of a direct object to complete the meaning of the sentence.
- The transitive verb does not requires the presence of a direct.
- Etc.

- The subject must precede the direct object:
  - If the subject is an attached pronoun and the direct object is an apparent subject.
  - If the subject and the direct object are an attached pronoun .
  - Etc.
- The direct object must precede the subject: :
  - If the direct object is an attached pronoun and the subject is an apparent subject.
  - If the relative pronoun to the direct object is in accord with the subject.
  - If the pronoun "إلا" or "انما" are located inthe right and in the left of the subject.

- The subject must succeed the verb
- Forever the verb still in the singular, the agreement can be only in kind and ever in number.

Fig. 1 Some rules in the Arabic Language

We illustrate in the Fig. 2 some properties through an Arabic sentence means "I walked in the beautiful street, then I entered to the store."

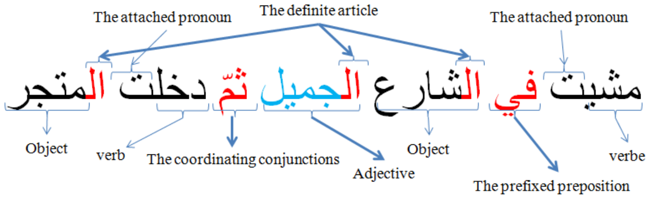


Fig. 2 Example of properties through two Arabic sentences.

The principles which form a grammatically valid sentence are "complementary and accomplice" inorder to achieve a clear expression and to ensure semantic consistency.

1) *The frequent errors in the Arabic language*

The processing difficulty lies in the ambiguity caused by the partial voyellation, agglutination and the free order of lexical tokens in the sentence [5]. The two major problems of automatic processing of Arabic language are: Agglutination of words and the lack of vowels in writing. In fact, unlike Latin languages, Arabic is an agglutinative language. It is composed of lexical tokens bonded to each other and carrying several morphosyntactic information. Example: personal pronouns can be attached to nouns (نصحه = he advised him).

The Arabic language is also specified by Voyellation which is necessary to the correct reading and understanding of a text. It allows to distinguish lexical units that have the same representation; Example: ذَهَبَ: to go;ذَهَبٌ : noun which means gold; كَتَبَ : to write; كُتُبٌ : noun which means books.

All these features make the risk of committing a hidden error greater than for other languages especially Latin. We can divide the errors committed by surfers into four categories: Syntactic errors, Semantic errors, Structural errors, Pragmatic errors.

A syntactic error affects the phrase structure such as words order, dependencies, and agreement [5].

إكتسب أموالا كبيرة الولد

إكتسب = verb « فعل»; الولد= subject « فاعل» ;
أموالا كبير= attribute « مفعول به» ;

In a verbal phrase, the subject must precede the attribute which is not the case in our example.

According to [8], Semantic errors relate to every word resembles typographically like the correct word, but which is semantically invalid in its context.

إكتسب الولد أموالا كثيرة (كبيرة)

The boy won much (big) money

The substitution of the character ب by the character ث caused a semantic incoherence. The adjective "كبيرة" (big) is used instead of "كثيرة" (much), so the sentence became semantically incorrect.

Structural errors: this type of errors violates the essential coherence relations in a text, such as an enumeration violation [9].

احتاج إلى ثلاثة مكونات: سكر و حليب

I need three components: sugar and milk

Pragmatic errors: present some anomaly according to the goals and plans of the discourse participants. [10] made a statistic on the English language which showed that 40% of 4218 errors are errors producing words which are lexically valid but semantically invalid.

Indeed, although this type of error preserves the syntactic validity of the sentence where it is, it is mostly breaking its semantic consistency, which makes the sentence incomprehensible by a human. According to [11], this type of errors presents 25% of spelling errors in a reference corpus.

Problems concerning the semantic consistency are a challenge to overcome. The difficulty of these problems is greatly amplified by the complexity of the Arabic language. So, we can conclude that the difficulty to develop an error detection system is principally resulted from the nature of the frequent errors.

III. ONTOLOGY AND ERROR DETECTION

A. *The errors detcetion approches*

Detecting simple errors is now a standard feature of advanced text treatment, but many errors remain difficult to identify.To remedy this problem, several studies have been conducted in the Latin language. For example [12] have presented an approach that combines the Winnow algorithm which is based on the method of neural networks with weighted majority voting, using near and adjacent words as properties.

[13] suggested a solution for the detection and correction of ontologies incoherence, that is to say, the resolution of semantic contradictions. The approach is based on anti-patterns wich define bad modelling choice causing errors. [15], two Chinese researchers have proposed a novel and efficient algorithm for the system of Chinese spelling error correction.

[9] proposed a method based on the trigrams to detect errors in the English language. This method assumes that all trigram of words in the text that exists in the British National Corpus is correct, and all trigram missing is a probable error. [14] presented a new evaluation algorithm of [9] such that the obtained results can be compared with other methods, then built and evaluated some variants of the algorithm using windows with fixed size.

[15] proposed two methods of post-processing to correct Arabic words issue from OCR systems. [16] have created an hybrid method of analysis to detect a hidden semantic errors in the Arabic text. They combined a several statistical or mixed methods which represent each word according to the near and far context which it appears, and compares this performance to previous performances obtained during learning.

[17] proposed an approach for spelling error detection, based on two methods, the direct method, to match words in an input text against a dictionary, and a character-based language modelling method in case such a word list is not available. [18] used an hybrid model to spell checking Arabic word based on morphological structure of Arabic word. He used a consistent root-pattern relationships and a morphographemic rules to specify the word recognition.

B. *Ontology*

1) *Definition*

This use of ontologies allows solving different problems such as conception and indexing databases, integration and data sharing…

Since 1990s, the notion of ontology has rapidly spread in diversity research areas in computer science [19]. The first proposed definition of ontology in computer science was that of Gruber [20]: “ontology is an explicit specification of a conceptualization”. The notion of “conceptualization” was defined as follows : “the objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold them” [21].

In computer science, ontology is essentially a graph-based knowledge representation in which each node corresponds to a concept and each edge specifies a relation between two concepts [22]. Concepts can envelop a diversity of real world entities and abstractions such as objects, names, ideas, events, and types. Relations can enclose different unidirectional or bidirectional associations, containing connectivity, hierarchy, membership, functional mapping, and causation.

As computational models, ontologies have been widely used in artificial intelligence, natural language processing, and Web sciences. Ontologies will doubtlessly be indispensable in developing infrastructures for knowledge assisted visualization.

2) *Construction of ontology*

Four types of ontologies are frequently distinguished [23]:

- The Top-level ontology describes general concepts like space, actions, time, material objects, events, etc. These concepts are not dependent on a problem or a particular domain.
- Domain ontology are related to a particular discourse, they describe the existing knowledge corresponding to this universe.
- Process ontology describes generic tasks, such as banking.
- Application Ontology is the combination between Domain ontology and Process ontology.

In the literature, we found two approaches adopted for the construction of ontology. The first one consists on building a corpus and identifying terms and lexical relations. Then it is necessary to define the concepts and semantic relationships in a semi-formal language through semantic standardization. Concepts should be integrated and formalized concepts in a formal knowledge base [24]. The second approach presented by [25] consists in identifying relevant terms and synonymy relations between them.

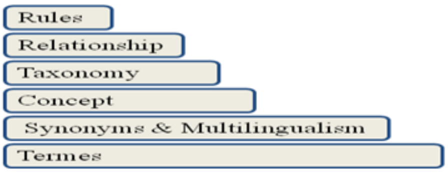


Fig. 4 Construction of ontology

We must realize the identification of concepts and their attributes as well as taxonomic relationships and not taxonomic that connect them. Finally, we must identify the rules specifying constraints on the properties of concepts and relationships (Fig. 4) [19].

C. *Formalisms for representing the ontology:*

Knowledge representation is manifested through several formalisms: We will mention two of the most used: the conceptual graphs and the description logics.

In computer science, ontology is essentially a graph-based knowledge representation in which each node corresponds to a concept and each edge specifies a relation between two concepts [22]. Relations can enclose different unidirectional or bidirectional associations, containing connectivity, hierarchy, membership, functional mapping, and causation. The conceptual graph model consists of two parts: a terminological part and an assertional part. The terminological part relates to the concepts, relationships and instances. The assertional part is for the representation of the domain knowledge assertions [21].

Several models of ontology representation language that we can call language have been developed. We mention the RDF model (Resource Description Framework) which is developed by the W3C (the World Wide Web Consortium).The RDF represents data from an XML vocabulary. In fact, it represents resources and relationships between resources through an independent description of applications via a triplet representation<subject, predicate, object>. The Subject present the resources having an Universal Ressource Identifier (URI). The Object consiste on the value of the predicate and can be a resource or a literal. Finaly the predicte is composted of relationship and resource between the subject and the object.

The RDF model can be represented as an oriented labeled graph. It is formalized via a triplet representation: the tops of the graph relate to the subjects and the objects and the triplets are represented by arcs (from subjects to objects). The RDF model can not model the semantics of an application model. However, the RDFS (Resource Description Framework Schema) model tries to close this loophole and has advantages over the RDF model [26].

Indeed, the RDFS data model provides a representation through a set of simple primitives for the development of the knowledge field in the form of classes and sub classes. Although this method has significant advantages and presents a strong improvement of the RDF model, it is still limited as an expressive model of the application domain [27].

OWL is a W3C standard; it is based on the RDF / XML syntax. OWL provides a description of the characteristics of the relationships between classes using tools such as: equity, cardinality, symmetry, transitivity, disjunction, etc. This is thanks to a richer vocabulary and formal semantics. In fact, OWL is a language which gives the possibility to a decidable reasoning but does not provide at the same time RDFS compatibility, which resulted in three versions up ward expressiveness: OWL Lite, OWL DL et OWL Full [28].

IV. Contribution

A. *Motivation*

Error detection is a complex task that can affect the semantic coherence of the sentence (section B) especially for the Arabic language which is characterized by its diversity and morphological richness.

The idea is based on the fact that the error is an incorrect modified rule; the ontology of grammatical rules will explain the hidden meaning of the rules and will allow us to reason in the extended semantic grammar of the Arabic language. Based on this reasoning, we can locate the error. In our work we will focus on the errors syntactically correct but that deform the meaning of the sentence. This type of error can not be detected by a simple system.

In our approach, we exploit the richness of the morpho-syntactic connections of Arabic to locate errors. In Fact, Arabic is based on properties from which we can build a reasoning to detect mistakes. The complexity of the problem consists in interpreting wrong grammatical rules. The ontology allows us to achieve perfectly our objective and support easily the complexity of the problem. Through the ontology representation we can illustrate the grammatical rules. Therefore, we can conclude wrong rules based on those interpretations So the objective is to identify the error through the grammatical rules by using ontology.

B. *Adopted Method*

In our approach we model the grammatical rules in the form of contexts and roles. All illustrated rules are for verbal phrase (جملة فعلية). The entities identified on the sentence are represented by concepts through the nodes such as (Lézem = لازم, Moutaadi = متعدي, etc). Arcs represent relations between the different concepts such as: Consists of: متكون من , Followed by: يليه , Type: نوعه

In what follows, we will present explicitly some grammatical rules that characterize the Arabic language before their formulation in the conceptual map:

- The adjective inherits fours character from the apparent noun: a common or a definite noun, masculine or a feminine noun, the Inflection, the number (singular, dual, plural).
- The pattern Mafeal مَفْعَل (bureau مكْتب) and Mafeil مفعِل (position موقع), give reference to the places or time
- The pattern of the intransitive verb Faoul فَعُل, Ifalalla افعَلَّ, Ifanlalla افْعنْلَلَ , faiila فَعِـل , infaala انفعّل (قوِي, اقشعر , قَصُرَ , احْرَنجم)
- If the pattern of adjective is :فعول , فعيل (صبور,عجوز) the adjective should still masculine even if the apparent noun is feminine.

According to Fig. 5, we can generate an expression based on concepts and roles, in referring to the conceptual graph and the order of terms constituting the sample.

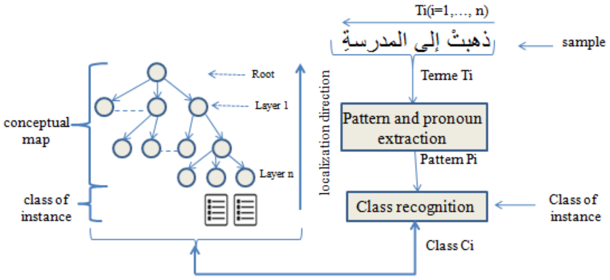


Fig. 5 Process of extraction rules

To do this we must recognize the concepts of each term and the roles that refer to him as follows:

- Extract the pattern or the root of the term Ti, the personal pronouns, attached pronouns and the definite article.
- Recognize the class of instances based on the pattern or the root of the terme Ti.
- Remount in the graph to locate the father concept in layer 1. In this step we must refer to the order of the term in the sample to avoid the ambiguity and to choose the right path to the father concept on layer 1.
- Get all relationships that are related to the father concept
- Take into account the concepts and roles obtained for the terms T_j such that j <i and i <1 in order to identify the father concept of T_i.

In the following we will present the rules of the correct sentence and the rules of error through this example:

إكتسب ولدٌ أموالا كثيرة

For a correct sentence the extracted rules are presented as follow:

Verb type intransitive ^ verb followed_by subject ^ subject type apparent subject ^ subject temined_by "TANWINE_THAMA" ^ subject followed_by object^ object type apparent subject ^ object type countable ^ objet contains adjective ^ adjective type countable

If the test sentence contains a semantic error example:

إكتسب ولَدَ أموالا كبيرة

The incorrect rules will be represented as follow:

Verb type intransitive ^ verb followed_by subject ^ subject type apparent subject ^ subject termined_by « FATHA» ^ subject folowed by object ^ object type apparent subject ^ objcet type countable ^ object contains adjective ^ adjective type quantifiable

By referring to the map we find that (subject termined_by « FATHA ») and (adjective type quantifiable) don't exist, then we have the answer in relationship as follows

- ˥(subject termined_by « FATHA») → (sujet ˥termined_by « FATHA»).
- ˥(adjective type quantifiable) → (adjectif ˥type quantifiable)

The equivalent of the relation ˥termined_by and ˥type exist in a dictionary that contains an opposite direction of each relation: ˥termined_by→ does not terminate ; ˥ type→ is not a type

The difficulty in our work is to enrich the ontology of grammatical rules; because to identify the extended semantics of probable errors we need a very rich map of rules.

V. Conclusion

In this article, we introduced an approach which is based on the formalisms of the ontology. In fact, the interest of this approach is to take advantage of reasoning on the conceptual graph in order to model explicitly the mistakes and locate in them in the sentence. The detection of the semantic mistake is based on modeling the grammatical rule through the concepts and the rules to make a reasoning on graph and come closer to the mistake until the localization; except that the efficiency of the approach depends enormously on the volume of the grammatical rule's corpus, the more the graph includes rules, the more the result of the detection is refined.

References

[1] M. Paul Lewis, “Ethnologue: languages of theworld”. Dallas, Texas: SIL International, pp. 5, 2013.

[2] W. Chung, “ Web searching in amultilingual worl”, Magazine:Commun. ACM- Web searching in a multilingual, vol. 51 Is 5,pp. 32–40, 2008.

[3] N. Habash, “Introduction to Arabic natural language Processing. Synthesis Lectures on Human Language Technologies”, Morgan &Claypool Publishers , pp. (15–17, 21, 27,29, 56), 2010.

[4] W. Black, S. Elkateb, P. Vossen, “*Introducing the arabic wordNet project*”, In Proceedings of the third International WordNet Conference (GWC-06), 2006.

[5] S. Gahbiche-Braham, “Improvements for machine translation systems using linguistic and thematic analysis : an application to the translation from arabic”, University Paris Sud - Paris XI, septembre 2013.

[6] M. Al-Kabi, A. Nawaf, M. Al-Ayyou, “An analytical study of arabic sentiments: Maktoob case study”, Internet Technology and Secured Transactions (ICITST), 8th International Conference for , vol., no., pp. 89-94, 2013

[7] A. Ibrahim, and T. Elghazaly, “Improve the automatic summarization of arabic text depending on Rhetorical structure theory”, Artificial Intelligence (MICAI), 12th Mexican International Conference on , vol., no.,pp. 223-227, 2013.

[8] C. Zribi, H. Mejri and B. Mohamed, “Un analyseur hybride pour la détection et la correction des erreurs cachées sémantiques en langue arabe”. Traitement Automatique des Langues Naturelles : à Toulouse. 2007.

[9] S. Verberne, “Context-sensitive spell checking based on word trigram probabilities”. Master thesis Taal, Spraak & Informatica University of Nijmegen, 2002.

[10] R. Mitton, “Spelling checkers,spelling correctors and the misspellings of poor spellers”, Information Processing and Management: an International Journal, Vol. 23 n° 5, pp. 495-505, 1987.

[11] C.W. Young, C.M. Eastman and R.L. Oakman, “An analyses of ill-formed input in natural language queries to document retrieval systems”, In Information Processing and Management, Vol. 27, pp. 615-622, 1991.

[12] A.R. Golding, and D. Roth, “A winnow-based approach to context-sensitive spelling correction”, Machine Learning - Special issue on natural language learning. Vol. 34 Is 1-3, pp. 107 - 130, 1999.

[13] C. Roussey, F. Scharffe,O. Corchoand O. Zamazal, “Une méthode de débogage d’ontologies OWLbasées sur la détection d’anti-patrons”,21èmes Journées Francophones d'Ingénierie des Connaissances, Nîmes : France, 2010.

[14] W. Xiaolong and L. Jianhua, “Combine trigram and automatic weight distribution in Chinesespelling error correction”, Journal of computer Science and Technology, Vol. 17 Is 6, Province, China, 2001.

[15] S. Toufik and M. SELLAMI, “Deux méthodes morpho-lexicales pour la correction des mots Arabes issus des systèmes OCR”, African conference on Research in Data processing and mathematics applied CARI 2, 2002.

[16] C. Zribi and B. Mohamed, “Detection of semantic errors in Arabic texts”, Journal Artticial Intelligence, Elsevier Science Publishers, pp. 249-264, 2013.

[17] M. Attia et al, “*Improved spelling error detection and correction for arabic*”,pp. 103-112, *2012.*

[18] B. Haddad and Y.Mustafa, “Detection and correction of non-words in arabic: A hybrid approach” ,International Journal of Computer Processing of Languages ,pp. 237-257, 2007.

[19] H. Teguiak, “Construction d’ontologies à partir de textes : une approche basée sur les transformations de modèles”, Ecole Doctorale de Sciences et Ingénierie pour l'Information, Mathématiques: Informatique et Applications, 2012.

[20] J. Grant, W. Litwin, N. Roussopoulos and T. Sellis, “Query languages for relational multidatabases”. The VLDB Journal,The International Journal on Very Large Data Bases Vol. 2 Is 2, pp. 153–172, 1993.

[21] S. Ayad, “Une plateforme pour la construction d’ontologie en arabe : Extraction des termes et des relations à partir de textes”, Badji Mokhtar-Annaba University, 2013.

[22] S. Carpendale et al, “Ontologies in biological data visualization”. Computer Graphics and Applications, IEEE , vol. 34, no.2, pp. 8,15, 2014.

[23] N. Guarino, “Understanding, building, and using ontologies : A commentary to using explicit ontologies in kbs development”, International Journal of Human and Computer Studies, pp. 293–310, 1997.

[24] B. Biebow and S. Szulman, “Terminae: a linguistic based tool for the building of domain ontology”, In Proceedings of the 11th European Workshopon Knowledge Acquisition, Modeling and Management,EKAW ’99, London, UK. Springer-Verlag, pp. 49–66, 1999.

[25] B. Paul and C.B.M. Philipp, “Ontology learning from Text:An overview”, Vol. 123, chapter-.IOS Press, 2005.

[26] C. Amina, “Gestion des dépendances et desinteractions entre Ontologies et Règles Métier”, University Pierres et Marie Curi- Paris VI, 2013.

[27] M. Radja,“Ontologies et services aux patients :Application à la reformulation des requêtes”, Ecole Doctorale EDISCE, Université Joseph Fourier-Grenoble I, 2009.

[28] C. Chedlia, “Contribution à la définition d’une méthode de conception de bases de données à base ontologique”, Ecole Doctorale : Sciences et Ingénierie pour l'Information, Mathématiques, 2013.