Arabic Language Textual Content Research Overview with Scope Interrelations Modeling

Heba. Konswah

***Abstract*—**Arabic Language research covered many areas of textual content annotation and information retrieval. This paper, introduces a model for interrelations between these areas and how may this help integrating and improving Arabic language text processing research. It will be concerned about Annotation, Named Entity Recognition, Ontology, Morphology, Search, Translation, and Classification related work current state and techniques.

***Keywords*—** Arabic Text, Annotation, Information Retrieval, Ontology.

I. Introduction

ARABIC text processing research is becoming very important to many applications, especially web based ones. It can aid in improving Arabic language text annotation and information retrieval.

Recently, Arabic language research started to cover many linguistic aspects; such as textual content annotation using ontology in different domains employing various techniques. The problem is each research was conducted in separation from other researches so they all lack integration, in some cases one domain may have more than one corresponding ontology.

In this paper, we introduced a precise review for most of the research that was conducted on Arabic language text, with highlighting the interrelations between research projects that have common work area and serving the same target. This will give the future researchers an integrated overview about the whole picture for Arabic language research state to enable them benefit from this in building their contribution in a cumulative way unifying Arabic language research findings into one comprehensive framework.

II. Arabic Language Text Processing Categories

A. Proposed Model for Arabic Language Research Intersections

A proposed model is introduced to visualize the intersection between various research scopes. As shown in Fig. 1 this model shows seven working areas: Annotation, Named Entity Recognition (NER), Ontology, Ontology Construction, Search, Translation, Morphology and Classification.

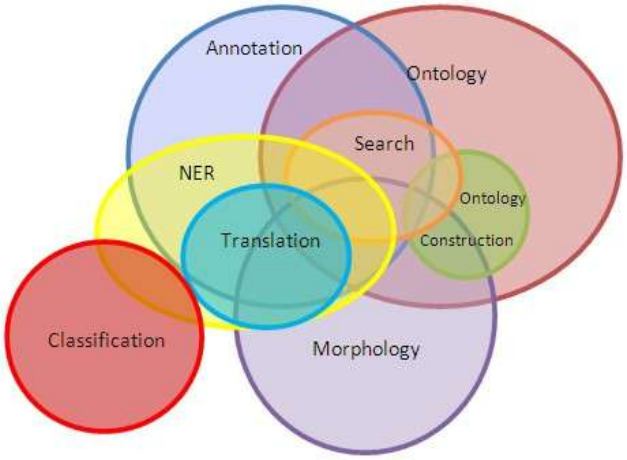

Fig. 1 Modeling Arabic Language Research Intersections

B. Arabic Language Research Categorized Findings

Each research scope intersected with another to produce a solution that is considered an added value in this scope. A list of work conducted in each covered intersection will be clarified. In each research, we will view the findings and techniques used.

1) Ontology Construction

As an example for ontology developed specifically for Arabic language is Al-Azhary ontology [1]. It is a lexical ontology that contains 26,195 words organized in 13,328 synsets, and records association relation between words. Its primary use is in automatic text analysis and Artificial Intelligence applications. It is based on the Holy Quran which has 77,439 words as seeds, stored without repetition with IExcelAPI. Relation building was fully manual process using Arabic/Arabic dictionaries. Ontology creation and search GUI were developed with Apache Jena.

Arabic language ontology was developed by Jarrar [2]. It is conducted with semantic relations between Arabic concepts. It was built through mining Arabic concepts from dictionaries then mapping these concepts to Word Net automatically with inherited relations. A link between all concepts to Arabic core ontology with reformulations is established according to ontology guidelines. Mazari et al. [3] research uses statistical methods; Repeated Segments to identify the domain relevant

Heba. Konswah is with Faculty of Commerce, Alexandria University, Alexandria, Egypt (phone:+201065835237 ; e-mail: h.saied89@gmail.com).

terns retrieved form GOLD (General Ontology for Linguistic Descriptions) and Co-Occurrences to update the ontology by linking the extracted concepts to the ontology. Unlike the previous researches Elayeb et al. [4] introduced an aiding approach for handling domain related terms disambiguation whilst building the ontology. It uses shallow morphosyntactic parser and applies qualitative approach to weigh term in the document according to their frequencies. This experiment was only applied on three domains (Drink, Marriage and Purification).

2) Ontology Construction-Morphology

A project for modeling Arabic language morphological rules was developed by Aliane et al.[5]. It focuses on Arabic traditional grammar. First, linguistic concepts were chosen manually from Arabic linguistics and relating them to GOLD concepts formulating the core ontology then repeated segments calculations are used to extract new concepts to enrich the ontology.

3) Ontology Search

A semantic search engine was developed by Ishkewy and Harb [6] using SPARQL and OWL to aid non-Muslims in reaching accurate information from Holy Quran. It is based on Islamic ontology and Al-Azhary lexical ontology. A similar research was conducted by Khan et al. [7] to search the Holy Quran using SPARQL and Protégé. It differs than the previous research that it is based on domain ontology about creatures mentioned in the Holy Quran.

An Arabic semantic search engine was developed by Moawad et al. [8] based on a very rich vocabulary Arabic ontology of concepts' attributes, inheritance relations and association relations in the computer domain. The user query is analyzed with semantic query analyzer to find related concepts then a semantic ranker shows the most relevant documents first. Interfaces are developed to couple the module with syntactical search engines.

4) Ontology-Translation-Search

Abdelali et al. [9] introduced a cross language information retrieval system. This system uses keizai which is a cross language interactive retrieval and summarization system. It takes a query in English language and translates it using English lexicon ontology with mapping to Chinese and Arabic concepts then the most relevant documents are retrieved.

5) Annotation

NEMLAR project (Network for Euro-Mediterranean Language Resources) in [10]. It is project for building a network of specialized partners, to support in the development of Arabic language resources. It was built on a corpus of annotated written Arabic that contains more than 500K word, Arabic speech recordings corpus of more than 10 hours of read speech and Arabic broadcast news speech corpus of about 40 hours. Miscellaneous domains were covered in this corpus. Types of annotation applied include Arabic lexical analysis, part-of-speech (POS) tagging, and phonetic transcription (Vowelization).

A research that was made by Alrahabi et al. [11] presented a platform for automatic annotation of semantic relations of quotation segments in Arabic and French based on contextual exploration. EXCOM (Exploration Contextual Multilingual System) is a rule based system that uses a semantic map of enunciative modalities in direct reported speech. The rules are used to split paragraph sentences then concept exploration rules are examined if satisfied, the engine attributes the annotations or calls another rule recursively.

6) Annotation-Ontology

An automatic annotation tool was presented by Al-Bukhitan et al. [12] that support the semantic annotation of Arabic language web documents. AMAST (Automatic Arabic Semantic Annotation Tool) a prototype that was developed with documents of three domains (food, nutrition, and health) for evaluation. The documents were manually annotated according to domain ontology and compared to the prototype annotations.

7) Annotation-Ontology construction

Reference [13] describes the design and implementation of a lexical ontology of for Arabic semantic relations that facilitates semantic annotation of the Arabic textual content. Semantic ontology was developed using protégé, with major entities of linguistic units, semantic fields, and semantic relations. The ontology was evaluated using a system prototype that was developed to perform ontology based semantic annotation including entity identification and relation extraction.

8) Annotation-Morphology

Annotating Arabic text using statistical approach was presented by [14]. It uses Hidden Markov Model (HMM) to represent the internal structure of the Arabic sentence to consider the logical linguistic sequencing in the POS tagging process. Morphological analysis was used to reduce the lexicon tags size as the principle aspects of Arabic grammar classification to develop the tagging scheme in an expandable form. It was also used in segmenting Arabic words into their prefixes, stems, and suffixes. A corpus of old text of third century Hijri was created and tagged manually to train and test the tagger.

In the Quranic domain Sawalha et al. [15] produced morphological analyzer and POS tagger Fine-Grain. The lemmatizer program extracts the root of a word then the word decomposed into proclitics, prefixes, roots, suffixes, and eclitics then the analyzer adds the appropriate linguistic tags according to the morphological features tag set. Dukes and Habash [16] introduced Quranic morphological annotator that gives a set of morphological annotation for each word in the unique language of the Holy Quran. Initial automatic tagging was done using BAMA (Buck Walter Arabic Morphological Analyzer) then one-to-one word mapping was established with manual verification in cases of required disambiguation. The corpus was reviewed by a trained Arabic linguistic

annotator and corrections were made, the corpus also was put online for community volunteer corrections with approvals.

9) Annotation-NER

Alanazi et al. [17] presented NAMERAMA a named entity recognition system that was developed to extract Names Entities (NEs) in the medical domain such as diseases' names, symptoms, treatment methods, and diagnoses. The system underlying technology is Bayesian Belief Networks (BBN), it uses patterns that contain some verbs related to certain entities to assist in the investigation of the context of named entity in the medical domain language structure.

A Natural Language Processing (NLP) project was developed by Stanford University [18] it has a set of language analysis tools. Stanford Arabic core NLP system can identify NEs like (Persons, Organizations, and Locations) and co references as it gives words forms, their point of speech and markup the structure of sentences in terms of phrases and words dependencies. Stanford Parser can work out grammatical structure of sentences, subject, object of a verb based on Penn Arabic Tree Bank.

A toolkit developed by Diab [19] is AMIRA which can process Modern Standard Arabic language. It provides a clitic tokenizer, POS tagger, and base phrase chunker that can identify noun phrases and verbal phrases components as a first step towards shallow syntactic parser. It is based on supervised learning with underlying technology of Support Vector Machines (SVM) without knowledge of deep morphology. Another natural language annotation tool developed by Morton et al. [20], Word Freak which can be used to annotate English, Chinese, and Arabic text. It provides automatic content extraction of NEs, co reference relations, and active learning of rapid annotation of new text. Word Freak is a Java coded tool that is integrated with automatic annotators including sentences detectors, POS tagger, Parsers, and co reference revolvers.

10) Annotation-NER- Ontology -Translation-Search

To improve the Question/Answering systems Abouenour in [21] investigated the usefulness of enriching the Arabic Word Net (AWN) NEs by using Yago ontology. Yago ontology covers more than 2 million NEs and it was built from AWN and Wikipedia it was also connected with SUMO (Standard Upper Merged Ontology). The proposed system extracts the NEs in Arabic from each question and translates them into English using GTA (Google Translation API) as it solves disambiguation. The system extracts the Yago relevant English entities and facts then the sub release of Yago related to the considered question is translated using GTA. Finally, a mapping between the NEs in Arabic Yago and their entries in AWN is established according to synonymy, hypernymy, hyponymy, and SUMO relations.

11) Annotation-NER-Ontology-Morphology-Search

An artificial intelligent approach was presented by Atwell et al. [22] for building Quranic knowledge map, a structured large-scale online resource to help in understanding the Quran. The proposed model contains Arabic morphological analyzer and syntactic parser to annotate Quranic text and enable morphological and concept topic map search. The model is based on a readable structured datasets of linguistic and semantic information that include syntactic tree bank, Quranic word net, and ontology of Quranic concepts.

12) Annotation-NER-Morphology

Benajiba et al. [23] introduced a study that investigates the impact of using different features tag set on NER using three discriminative machine learning frameworks, Support Vector Machines (SVM), Maximum Entropy (ME), and Conditional Random Fields (CRF). The approach first explores individual features (contextual and lexical) then it ranks them according to their impact and the top N-top (starting from N=1 to N=22) features are combined and evaluated using SVMs, ME, and CRFs.

13) NER-Translation

Alkhalifa et al. [24] focuses on extracting Arabic NEs from the Arabic Wikipedia (AWP) to extend the Arabic Word Net (AWN) NEs coverage. First, a dataset of geographical NEs is formed from three resources including GEONAMES database, Gazatteer of countries, and NEs recorded in NMSU (New Mexico State University), a manual validation is applied as these resources may have inconsistencies. The algorithm extracts English NEs from Princeton Word Net (PWN) then it links these entities to English Wikipedia (EWP) pages and "interwiki link" to their Arabic pages is looked for and the titles are returned as Arabic NEs.

14) NER-Classification

Al-Thubaity et al. [25] presents a project for Arabic text classification KACST (King Abdulaziz City for Science and Technology). It first forms a dataset that covers different text genres in the Arabic news domain collected form Linguistic Data Consortium (LDC), Arabic NEWSWIRE and Arabic Giga word corpus. Then the features which relate to the classification content were selected using Arabic Text Classification (ATC) tool that extracts lexical features and calculates the feature frequency profile to identify their importance and generate training and testing matrices with their selected features' weights. The classification algorithm of KACST is to experiment with all classification techniques provided by Rapid Miner 4.0 which are C5.0, Neural Network, SVM, and Naïve Bayes.

15) Morphology

A very large set of Arabic morphological rules was presented by Boudlal et al. [26]. Al-Khalil is a morpho-syntactic parser for standard Arabic words developed in Java. Al-Khalil Morpho Sys was made of integrated linguistic data resources including datasets of roots, vocalized patterns and proclitics and enclitics tables. The system produces a table of vocalized stem, grammatical category, and all the possible solutions associated with their morphosyntactic features.

TABLE I
RESEARCH RESULTS

Research	Results
Annotation	
Automatic Annotation of DRS[11]	Declaration of Speaker: 75% (recall), 71% (precision) and 73% (F-measure)
Web Documents Semantic Annotator[12]	Food: 72.4% (recall), 86.6% (precision) and 78.9% (F-measure), Nutrition: 65.8% (recall), 83.7% (precision) and 73.7% (F-measure), Health: 68.5% (recall), 84.0% (precision) and 75.5% (F-measure)
POS Tagger [14]	Tagging Accuracy 96%
Fine Grain Morphological Analyzer [15]	85% Correctly analyzed
Quran Morphological Annotation [16]	72% (recall), 83% (precision) and 77% (F-measure)
NER in the Medical Domain [17]	69.23% (recall), 72.97%(precision) and 71.05% (F-measure)
Using Yago Ontology [21]	Accuracy before using Yago 17.49%, accuracy after using Yago 23,53%
AMIRA Tool[19]	Tokenization (F-Measure) 99.2%, POS Tagging Accuracy 96%, and BPC (F1 Measure) 96.33%
Classification	
KACST Tool[25]	Classification Accuracy: C5.0 84.43%, SVM 76.10%, Naïve Bayes 75.66%, and Neural Networks 63.78%

III. DISCUSSION AND CONCLUSION

In our study, we have reviewed most of the research that was conducted on Arabic language textual content from many aspects. Annotation and Classification research results (which can be measured numerically) are shown in Table I, the results indicate a good level of accuracy.

The main finding of this study is that all the research projects lack integration with each other. As the ontology is considered a basic core of textual content work, we recommend that the Arabic language text research community make their platforms ontology based, which means according to the proposed model that all the circles should be included or at least intersected with the ontology circle. As a result; they can later integrate their domain ontologies into one main inclusive ontology, which covers all the mentioned domains with the ability to be easily extendable in the future.

REFERENCES

[1] H. Ishkewy, H. Harb, and H. Farahat, "Azhary: An Arabic Lexical Ontology," in *Proc. Int. Journal of Web & Semantic Technology*, vol. 5, October, 2014.

[2] M. Jarrar, "Building A Formal Arabic Ontology," in *Proc. The Experts Meeting on Arabic Ontologies and Semantic Networks*, Tunis, April, 2011.

[3] A. C. Mazari, H. Aliane, and Z. Alimazighi, "Automatic Construction of Ontology From Arabic Texts," in *Proc. Int. Conf. on Web and Information Technologies*, Sidi Bel-Abbes, ICWIT, Algeria, April, 2012.

[4] B. Elayeb, Y. Slimani, I. Bounhas, and F. Evrard, "Organizing Contextual Knoweldge For Arabic Text Disambiguation and Terminology Extraction," *Knowledge Organiztion*, vol. 38, no. 6, pp. 473-490, Jan. 2011.

[5] H. Aliane, Z. Alimazighi, and M. A. Cherif, "Al-Khalil: the Arabic Linguistic Ontology Project," in *Proc. Int. Conf. on LERC*, May, 2010.

[6] H. Ishkewy and H. Harb, "ISWSE: Islamic Semantic Web Search engine," *Int. Journal of Computer Applications*, vol. 112, no. 5, Feb, 2015.

[7] H. Khan, S. M. Saqlain, M. Shoaib, and M. Sher, "Ontology Based Semantic Search in the Holy Quran," *Int. Journal of Future Computer and Communication*, vol. 2, no. 6, Dec, 2013.

[8] I. F. Moawad, M. Abdeen, and M. M. Atef, "Ontology-Based Architecture for An Arabic Semantic Search Engine," in *Proc. 10th Conf. on Language engineering*, Dec, 2010.

[9] A. Abdelali, J. Cowie, D. Farwell, B. Ogden, and S. Helmreich, "Cross-Language Information Retrieval using Ontology," in *Proc. TALN2003*, Jun, 2003, pp. 72-86.

[10] N. Paulsson, S. Haamid, S. Krauwer, C. Bendahman, H. Fersoe, M. Rashwan, B. Haddad, C. Mukbel, A. Mouradi, A. Al-Kufaishi, M. shahin, N. Chenfour, and A. Ragheb, "Building Annotated Written and Spoken Arabic LRs in NEMLAR Project," in *Proc. LERC*, 2006, pp. 533-538.

[11] M. Alrahabi, J. Descles, "Automatic annotation of Direct Reported Speech in Arabic and French According to a Semantic Map of Enunciative Modalities," *Advances in Natural Language Processing*, Springer Berlin Heidelberg, pp. 40-51, 2008.

[12] S. Al-Bukhitan, T. Helmy, and M. Al-Mulhem, "Semantic Annotation Tool For Annotating Arabic Web Documents," *Procedia Computer Science*, vol. 32, pp. 429-436, Dec, 2014.

[13] M. Al-Yahya, M. Al-Shaman, N. Al-Otaiby, W. Al-Sultan, A. Al-Zahrani, and M. Al-Dalbahie, "Ontology-Based Semantic Annotation of Arabic Language Text," *Int. journal of Modern Education and Computer Science*, vol. 7, no. 7, p. 53, July, 2015.

[14] Y. O. ElHadj, I. A. Al-Sughayeir, and A. M. Al-Ansari, "Arabic Part-of-Speech tagging using the Sentence Structure," in *Proc. Int. Conf. on Arabic Language Resources and Tools*, Cairo, Egypt, Apr, 2009.

[15] M. Sawalha, E. Atwell, "Fine Grain Morphological Analyzer and part-of-Speech Tagger for Arabic Text," in *Proc. 7th Int. Conf. in Int. Language Resources and Evaluation*, LREC, May, 2010, pp. 1258-1265. *European Language Resources Association*, ELRA, 2010.

[16] K. Dukes and N. Habash, "Morphological Annotation of Quranic Arabic," in *Proc. LREC*, 2010.

[17] S. Alanazi, B. Sharp, and C. Stanier, "A Named Entity Recognition System Applied to Arabic Text in the Medical Domain," *Int. Journal of Computer Science Issues*, IJCSI, vol. 12, no. 3, p. 109, May, 2015.

[18] Stanford University: The Stabford Natural Language Processing Group. Available: http://nlp.stanford.edu/software/

[19] M. T. Diab, "Second Generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS Tagging, and Base Phrase Chunking," *in Proc. 2nd Int. Conf. on Arabic Language Resources and Tools*, 2009.

[20] T. Morton and J. Lacivita, "Word Freak: An Open tool for Linguistic annotation," in *Proc. of the North America chapter of the Association for Computational Linguistics on Human Language Technology: Demonstration*, vol. 4, May, 2003, pp. 17-18. Association for Computational Linguistics.

[21] L. Abouenour, K. Bouzoubaa, and P. Rosso, "Using the Yago Ontology as a Resource for the Enrichment of Named Entities in Arabic Word Net," in *Proc. Int. Conf. on Language Resources and Evaluation Workshop on Language Resources and Human Language Technology for semantic Languages*, LREC, 2010, pp. 27-31.

[22] E. Atwell, C. Brierley, K. Dukes, M. Swalha, and A. Sharaf, "An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet," in

Proc. 3rd National Information Technology Symposium, Leeds, NITS, 2011, pp. 1-8.

[23] Y. Benajiba, M. Diab, and P. Rosso, "Arabic Named Entity Recognition: A Feature-Driven Study," IEEE Trans. Audio, Speech and Language Processing, vol. 17, no. 5, pp. 926-934, Jul, 2009.

[24] M. Alkhalifa and H. Rodriguez, "Automatically Extending NE Coverage on Arabic Language Processing," in *Proc. 3rd Int. Conf. on Arabic Language Processing*, Rabat, CITALA, Morocco, May, 2009.

[25] A. Al-Thubaity, A. Al-Muhareb, S. Al-Harbi, A. Al-Rajeh, and M. Khorsheed, "KACST: Arabic Text Classifier Project Overview and Preliminary Results," 2008.

[26] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. O. A. O. Bebah, and M. Shoul, "Alkhalil Morpho Sys: A Morphosyntactic Analysis System for Arabic Text," in Proc. Int. Arab Conf. on Information Technology, 2010, pp. 1-6.