Arabic Meaning Extraction through Lexical Resources:

A General-Purpose Data Mining Model for Arabic Texts

Giuliano Lancioni, Ivana Pepe, Alessandra Silighini, Valeria Pettinari
Department of Foreign Languages, Literature and Civilizations, Roma Tre University
Rome, Italy
giuliano.lancioni@uniroma3.it, iva.pepe@stud.uniroma3.it, ale.silighini@gmail.com, val.pettinari2@stud.uniroma3.it

Ilaria Cicola
EPHE
Paris, France
ilaria.cicola@etu.ephe.fr
Department Italian Institute of Oriental Studies, Sapienza University of Rome
Rome, Italy

Leila Benassi, Marta Campanelli
Department Italian Institute of Oriental Studies, Sapienza University of Rome
Rome, Italy
benassileila@gmail.com, marta.campanelli@hotmail.it

***Abstract*— A general-purpose data mining model for Arabic texts (Arabic Meaning Extraction through Lexical Resources, ArMExLeR) is proposed which employs a chained pipeline of existing public domain and published lexical resources (Stanford Parser, WordNet, Arabic WordNet, SUMO, AraMorph, A Frequency Dictionary of Arabic) in order to extract a weakly hierarchised, single-predicate level, representation of meaning. This kind of model would be of high impact on the study of the computational analysis of Arabic for there is no such comparable tool for this language, and will be a challenge for the nature of its specificities. One should, in fact, cope with the unique writing system that is mostly consonant-based and doesn't always mark vowels explicitly. This is crucial when you want to analyze an Arabic corpus for the same consonantal ductus may be read in several ways.**

Keywords-Arabic data mining; content extraction; automatic parsing techniques; ontologies.

I. INTRODUCTION*

Data mining from Arabic texts presently suffers a series of shortcomings, some related to the specificities of Arabic texts and writing system [12, 13], some deriving from the scarcity, or plain lack, of lexical resources for Arabic analogous for what can be found for other languages [14].

Other tools routinely used as helpers in data mining cannot be successfully employed in analyzing Arabic texts as well, partly for these very reasons: e.g., statistical Machine Translation (MT) tools generally perform poorly for Arabic, both for the paucity of parallel texts and text memories and for the specificities of the language (the only other Semitic language with a reasonable amount of written texts available in electronic form, Modern Hebrew, for historical reasons has become closer to Indo-European languages in both lexicon and syntax) [15].

To overcome these difficulties, our project capitalizes on the use of existing Arabic lexical resources that are linked to larger, general-purpose resources, by devising specific strategies to fill the gaps in these resources. Resources are aligned through a pipeline which is fed by the input text and outputs, after several rings in the chain, a relatively hollow semantic representation that allows for further data mining operations, thanks to the Suggested Upper Merged Ontology (SUMO) format.

Next sections will discuss [II] the tools used in the project, [III] the workflow of the system, [IV] an example derivation, [V] test results and [VI] some conclusions and suggestions for further developments.

II. TOOLS

The ArMExLeR project employs a variety of tools. Some of them are shortly described in this section before tackling their role within the system.

A. Stanford Parser

The Stanford Parser is a statistical parser that is programmed in order to find the grammatical structure of the sentences. It analyses a text, parsing the phrases (constituency parser) and then finds out the Verb and then its Subject or Object (dependency parser). It is a probabilistic parser which uses knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences. Statistical parsers still make some mistakes, but their advantage is that they always give an answer that could be later corrected by a human. The lexicalized probabilistic parser implements a factored product model, with separate Probabilistic Context-Free Grammar (PCFG) phrase structure and lexical dependency experts, whose preferences are combined by efficient, exact inference, using an A* algorithm. Or the software can be used simply as an accurate unlexicalized stochastic context-free grammar

* All authors have contributed equally to this work , but since it refers to a modular project, Lancioni should be mainly credited for secs. 3 and 5, Pepe for sec. 2B, Silighini for sec. 4, Pettinari for sec. 2C, Cicola for secs. 1 and 2A, Benassi for sec. 6, Campanelli for sec. 2D.

 ISBN: 978-1-61208-311-7

parser and either of these yields a good performance statistical parsing system. As well as providing an English parser, the Stanford parser can be and has been adapted to work with other languages such as Chinese (based on the Chinese Treebank), German (based on the Negra corpus) and Arabic (based on the Penn Arabic Treebank). Finally, this parser has also been used for other languages, such as Italian, Bulgarian, and Portuguese.

Although the parser provides Stanford Dependencies output as well as phrase structure trees for English and Chinese, this component has to be implemented for Arabic. So, for now, we have an analysis of the sentences that cannot trace the subject or the object of a verb in a sentence, but we have a reliable parsing of constituents that could be used to deepen the analysis with other tools.

The Arabic Parser from Stanford can only cope with non-vocalized texts, the tokenization being based on the Arabic used in the Penn Arabic Treebank (ATB) and is based on a whitespace tokenizer. Segmentation is done based on the Buckwalter analyzer (morphology).

The character encoding is set on Universal Character Set (UCS) Transformation Format—8-bit (UTF-8), but it may be changed if needed. The normalization of the text is needed in order to analyze the text, because otherwise the parser cannot recognize the Arabic ductus, that is the representation of consonants and long vowels which are the only obligatory component of Arabic script, and often the only one actually present in texts (auxiliary graphemes, such as short vowels and consonant reduplication markers, must be deleted). For the part-of-speech (POS) tags, the parser uses an "augmented Bies" tag set that uses the Buckwalter morphological analyzer and links it to a subset of the POS tags used in the Penn English Treebank (sometimes with slightly different meanings) [1]. Phrasal categories are the same from the Penn ATB Guidelines [16]. As mentioned, there is no tool in the parser itself that can normalize or segment an Arabic ductus, so one is compelled to employ other tools in order to perform these tasks.

B. AraMorph

AraMorph [2] is a program designed to allow the morphological analysis for Arabic texts in order to segment Arabic words in prefixes, stems and suffixes according to the following rules:

- the prefix can be 0 to 4 characters in length;
- the stem can be 1 to infinite characters in length;
- the suffix can be 0 to 6 characters in length.

Each possible segmentation is verified by asking the software to check if the prefix, the stem and the suffix exist in the embedded dictionary. In fact, the program has three tables containing all Arabic prefixes, all Arabic stems and all Arabic suffixes, respectively. Indeed, if all three components are found in these tables, the program checks if their morphological categories are listed as compatible pairs in three tables (table AB for prefix and stem compatibility; table AC for prefix and suffix compatibility; table BC for stem and suffix compatibility). Finally, if all three pairs are found in their respective tables, the three components are defined suitable and the word is confirmed as valid.

Hereafter (Fig. 1), we put in evidence an example of an Arabic word and analyzed by AraMorph:

WORD NO. 10223: الثوري 17 occurrences

UNVOCALIZED TRANSCRIPTION: Al+vwry+
INPUT STRING: الثوري
SOLUTION: Al+vaworiy~+ ال*ثَوْرِيّ*
vaworiy~_1 [ثور]
ENGLISH GLOSS: the+revolutionary+

POS ANALYSIS:
Al/DET+vaworiy~/ADJ+

Figure 1. Excerpt of an AraMorph analysis of Meedan Memory

However, AraMorph presents some problems regarding the analysis of texts types that do not match to ideal text genre targeted by Buckwalter (newspaper texts and other Modern Standard Arabic non-literary texts). Indeed, the program shows three major weaknesses:

- it does not either fully or sparsely analyze vocalized texts;
- it does reject many words attested in some textual types which are not contained either in the sample of the text corpora chosen by Buckwalter, or in the lookup lists of AraMorph;
- there is neither any stylistic nor chronological information in the lookup lists; the same for a lot of transliterated foreign named entities which cannot be found in classical texts and in modern literary texts, giving rise to a number of false positives.

In order to overcome these problems, a group of researchers on linguistics at Roma Tre University had modified the original AraMorph in a new algorithm named "Revised AraMorph" (RAM), within a project of automatically analysis of ḥadīṯ corpora (SALAH project) [7]. The modifications, which are implemented to solve the weakness previously listed, are respectively:

- a mechanism which takes into account the vowels present in the text in order to reduce ambiguity linked to non-vocalized texts;
- a file with additional stems automatically extracted from Anthony Salomé Arabic-English dictionary (a work from the end of 19th century encoded in TEI-compliant XLM format) [11] and with additional lists of prefixes and suffixes with the relative combination tables of most frequent unrecognized tokens;
- a mechanism which removes automatically items in order to allow them matching to contemporary foreign named entities, especially proper names and place-names. In the other hand, the items above are not included in Salomé's dictionary (this way Arabic named entities which can be found in Classical texts are retained for the analysis).

 ISBN: 978-1-61208-311-7

C. A Frequency Dictionary of Arabic

Starting from the analysis of a 30-million-word corpus, Buckwalter and Parkinson's Frequency Dictionary of Arabic (FDA) [3] lists 5,000 most frequent Arabic words from Modern Standard Arabic as well as most important words from Egyptian, Levantine, Iraqi, Gulf and Algerian dialects. Each entry in the dictionary is organized as follows: headword and English translation(s), a sample sentence or context, English translation of the sample sentence or context and statistical information. The latter represents information about word dispersion figure and raw or absolute frequency, i.e., all the variants and inflected forms belonging to a specific lemma considered as an entry. The dictionary also provides important information about morphology, syntax, phonetics and orthography as well as usage restrictions and register variation.

Word ranking proceeds according to the value of a final adjusted frequency which is produced by multiplying word frequency and dispersion figure. Finally, the rank-order goes from the high-scoring lemma to the lower-scoring one.

An example of how an entry is generally arranged (information follows FDA's definitions) is in Fig. 2:

RANK FREQUENCY: 3835
HEADWORD: وَصْفَة
PART OF SPEECH: n.
ENGLISH EQUIVALENT: description, portrayal; (Medical) prescription; (Food) recipe
SAMPLE SENTENCE: كتب الطبيب المناوب لكل واحد منهم وصفة طبية
ENGLISH TRANSLATION: The doctor on duty wrote a medical prescription for each one of them
RANGE COUNT: 62
RAW FREQUENCY TOTAL: 434

Figure 2. Example of an FDA entry

D. Arabic WordNet

Arabic WordNet (AWN) is a lexical resource for Modern Standard Arabic based on the widely used Princeton WordNet (PWN) for English [5]. There is a straightforward mapping between word senses in Arabic and those in PWN, thus enabling translation to English on the lexical level. Each concept is also provided with a deep semantic underpinning, since, besides the standard Wordnet representation of senses, word meanings are defined according to SUMO.

However, AWN represents only a core lexicon of Arabic, since it has been built starting from a set of base concepts. In this sense, being the mapping with PWN relatively poor, this project uses in fact an AWN augmented model (AAWN) which extends this core WordNet downward to more specific concepts using lexical resources such as Arabic Wikipedia (AWp) and Arabic Wiktionary (AWk).

Wikipedia is by far the largest encyclopedia in existence with more than 4 million articles in its English version (English Wikipedia) contributed by thousands of volunteers and experimenting an exponential growing in size.

Arabic Wikipedia has over 224,000 articles. It is currently the 23rd largest edition of Wikipedia by article count and the first Semitic language to exceed 100,000 articles. The growing of Arabic Wikipedia is, however, very high so it seems that in a relatively short time the size of Arabic Wikipedia could correlate with the importance (of the number of speakers) of Arabic language. Wikipedia basic information unit is the "Article" (or "Page"). Articles are linked to other articles in the same language by means of "Article links". Wikipedia pages can contain "External links", that point to external URLs, and "Interwiki links", from an article to a presumably equivalent article in another language. There are in Wikipedia several types of special pages relevant to our work: "Redirect pages", i.e., short pages which often provide equivalent names for an entity, and "Disambiguation pages", i.e., pages with little content that links to multiple similarly named articles. A significant category of specific (non-ambiguous) concepts that can be drawn from Arabic Wikipedia in order to enrich AWN is that of Named Entities (locations, persons, organizations, etc.) that, once extracted from the mentioned resource, will be attached to existing Named Entities in PWN. In this operation, an important role is played by the "interwiki links" between Arabic and English Wikipedia, as shown in Fig. 3.

On the other hand, Wiktionary is a collaborative project to produce a free-content multilingual dictionary. It aims to describe all words of all languages using definitions and descriptions in English. It is available in 158 languages and in Simple English.

Designed as the lexical companion to Wikipedia, Wiktionary has grown beyond a standard dictionary and now includes a thesaurus, a rhyme guide, phrase books, language statistics and extensive appendices. It aims to include not only the definition of a word, but also enough information to really understand it. Thus etymologies, pronunciations, sample quotations, synonyms, antonyms and translations are included.

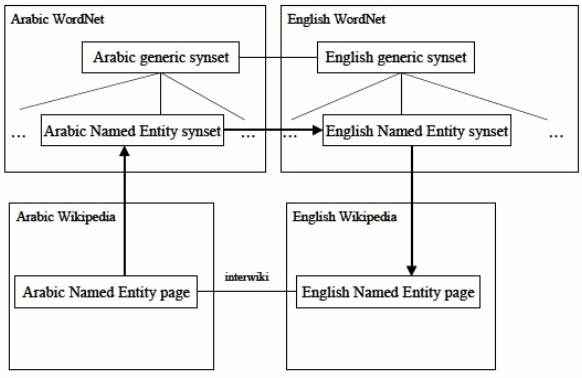

Figure 3. Relations between Arabic and English version of WordNet and Wikipedia

 ISBN: 978-1-61208-311-7

Wiktionary has semi-structured data. Its lexicographic data should be converted to machine-readable format in order to be used in natural language processing tasks.

Wiktionary data mining is a complex task. There are the following difficulties: the constant and frequent changes to data and schema, the heterogeneity in Wiktionary language edition schemas and the human-centric nature of a wiki.

III. WORKFLOW

Fig. 4 shows the general workflow of ArMExLeR.

The input to the system is Modern Standard Arabic written texts. The first analytical step is performed through the Stanford Word Segmenter (SWS) [4], which segments words into morphemes according to the ATB standard. SWS is not necessarily the best possible segmenter (e.g., it segments suffix pronouns, but not the article), but it is the best choice for the pipeline model, since it outputs an ATB-compliant segmentation, which is required by subsequent components.

The word-segmented input is submitted to the parsing component, Stanford Parser (SP), which statistically parses the input according to a factored model. Since, options for Arabic in SP are more limited than for English, it is not possible to get a dependency analysis, which would be more useful for content extraction. However, getting a standard parsing through the output of the (most probable) syntactic tree for an input sentence, is an invaluable contribution to a better semantic understanding of its element: e.g., identifying the subject and the object(s) of a (di)transitive verbs helps the system identify argument roles in relation to the verb - e.g., which argument is the agent and which the patient, - although linking of syntactic roles to argument roles is notoriously a nontrivial process.

An important role in the pipeline is played by the AraMorph component. The original AraMorph (AM) model is used by SP in order to lemmatize Arabic words; on the other hand, the RAM model [7] is used to select possible readings according to choices made by the syntactic parser.

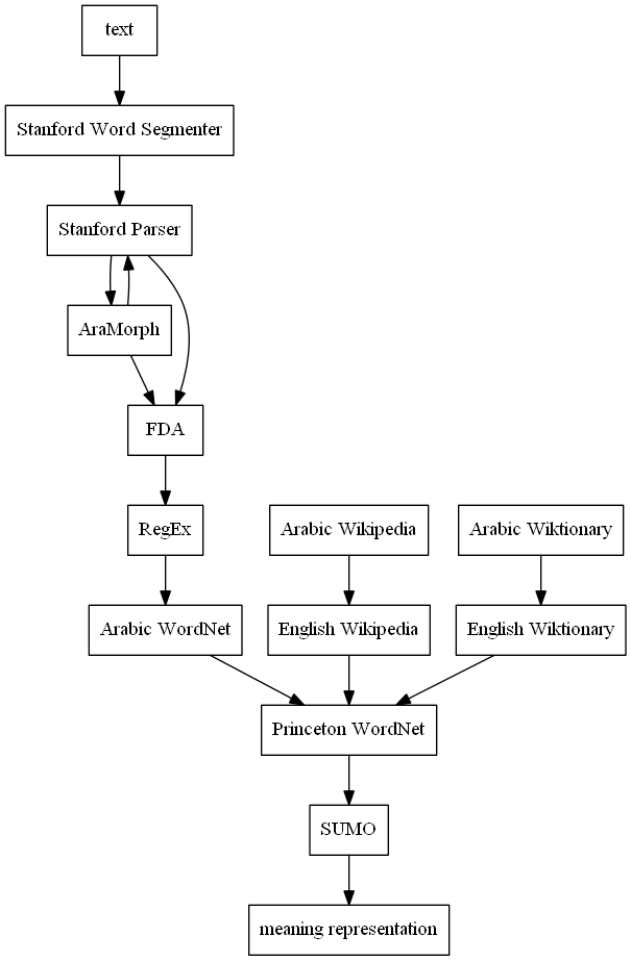

Figure 4. General workflow of ArMExLeR

In order to simplify the semantic linking task, in this step we worked on a subset of the analyses output by the parser, by filtering verb heads and the nominal heads of their arguments (plus possible introducing prepositions) through a regular expression component (RegEx).

The linking between RAM and FDA and between the latter and AWN realized by our research team is able to select the most probable reading and to link it to an AWN synset.

At this point, the system is fully within the semantic representation component: AWN is linked to the standard, Princeton WordNet 3.0 (PWN), which places the synset into a rich network of semantic relations. On its turn, PWN is entirely linked to SUMO, which allows the system to output a semantic representation in terms of ontologies, which can feed other components.

Since AWN is rather poor compared to the standard PWN, we use in fact an Augmented AWN model (AAWN) where AWN is supplemented by nonambiguous items drawn from AWp titles and sections, and AWk translations, linked to PWN by automatic linking [8, 9, 10]. This minimizes cases where a solution clearly exists, but it risks to be lost

 ISBN: 978-1-61208-311-7

owing to limitations in AWN (which was designed to represent a core lexicon of Arabic only).

IV. AN EXAMPLE DERIVATION PROCESS

The process can be better understood through an example. Let us start from one example (Fig. 5) drawn from the FDA corpus.

SENTENCE: نشرت الصحف قصيدة شوقى التى كتبها عن باريس بعد انتهاء الحرب الأولى

ENGLISH TRANSLATION: The newspapers published Shawqi's ode which he wrote about Paris after the end of World War I

Figure 5. Example sentence from the FDA corpus

The sample sentence is fed to SMS, which segments some of its graphic words into tokens (Fig. 6: tokens resulting from segmentation and other normalization steps are in boldface).

نشرت الصحف قصيدة شوقى التى **كتب ها** عن باريس بعد انتهاء الحرب **الاولى**

Figure 6. SMS tokenization of the sentence in Figure 5.

This segmented form of the text is input into the SP, which outputs (Fig. 7) a syntactic analysis.

The RegEx component extracts out of this syntactic tree the "core predications" (CPs: verb head and nominal heads of arguments), in order to simplify the generation of the semantic representation. This part of the system is clearly provisional, and it is likely to be widely improved in further development of the project. CPs extracted by the system are highlighted in light blue in the example.

The automatic WordNet-SUMO linking allows the system to immediately translate the PCs in terms of SUMO predicates.

Since SP has no dependency output available for Arabic — besides its general underperformance in dealing with Arabic texts compared to English ones,— such a parsing does not identify argument roles proper: e.g., in VSO clauses like the main clause in this example, we just have a sequence of NPs where nothing assures one of them is an agent, a patient, and so on. However, a general strategy that links roles output by this step to roles in entries for the relevant verbal concept in SUMO feeds back this step by assigning roles from the last to the first argument (owing to the general null subject property of Arabic).

CPs extracted by the RegEx component are linked to WordNet synsets through FDA (which selects the most frequent lemma in case of multiple possibilities) and AAWN. Synsets detected in the example are listed in Fig. 8.

نشرت = $publish_V2$(
الصحف = ,
قصيدة = $poem_n1$).

Figure 7. AAWN sysnets for an example entry.

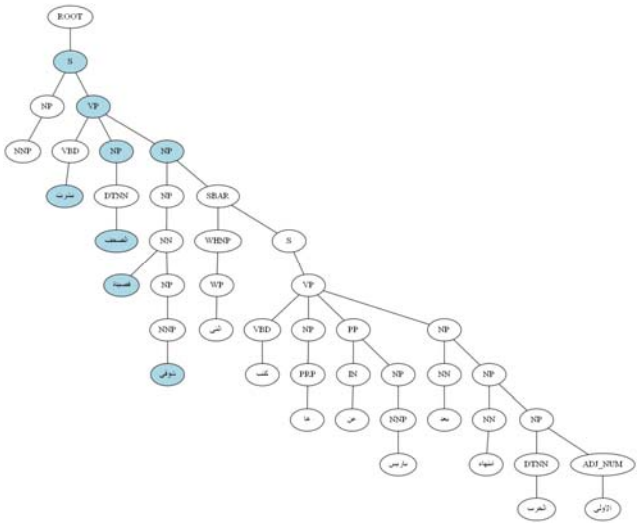

Figure 8. SP analysis of the sentence in Figure 5.

In this case, the result is (Fig. 9):

Publication (Corporation, Text).

Figure 9. SUMO representation for the example in Figure 7.

That is, the finer semantic relation has been transformed, and generalized, into a relation between SUMO concepts, which is expected to produce a better performance in data extraction.

V. RESULTS

Since the system relies on a complex pipeline of components which are only partially under the control of our research team, it is difficult to establish the best evaluation strategy for the results of the project. We decided to separate the output of the segmenter and parser components —which are taken "out of the box" from SWS and, respectively, SP— from the output of other parts of the system.

First, the ArMExLeR has been run against the whole Meedan translation memory (~20,000 Arabic-English sentence couples downloadable from [17]). Running performances are relatively slow since optimization has not been a core concern in this stage of the project yet.

Then, 350 analyses have been randomly extracted and assigned to two different members of our research team each for a single run of evaluation. While the test corpus might seem small, the relative homogeneity of the Meedan translation memory makes it large enough for our purposes, without requiring too many resources during the testing stage.

In 96 cases (27.4%), the SWS/SP output was discarded because it was judged significantly wrong (e.g., because the main verb had been misinterpreted as a noun) by both evaluators. Of the remaining 254 cases, a further 58 (16.6%) were discarded because they did not contain any predication

 ISBN: 978-1-61208-311-7

without anaphoric elements (which are not in the scope of the current model).

The evaluation of the system was performed on the remaining 196 cases (56% of the original sample). The analysis was deemed correct if the verb and at least one other PCs were regarded as properly assigned by at least one of the evaluators. This choice was motivated by the inherent problem in role-assignment caused by the lack of a dependency module for Arabic in SP (while such a module is available for English): therefore, the list of arguments and their relative order is not yet reliable. Only one agreement was deemed sufficient because a relatively high degree of disagreement between annotators has always been noticed for WordNet-related semantic projects (such as SemCor).

Results are summarized in Table I:

TABLE I. RESULT SUMMARY

error rate	60.54%
precision	64.90%
recall	74.56%
F measure	1.39

Comparing these results with other systems is not easy, since the ArMExLeR system evaluation applies to a specific subset of relations at the end of a relatively complex, automatic pipeline, which is not the case for other systems in Arabic text data mining. Therefore, we shall defer cross-comparison of our system to further research.

VI. CONCLUSIONS AND FURTHER DEVELOPMENTS

The ArMExLeR project shows a number of interesting features, which pave the way to further refinements and developments.

First, the system performs reasonably well, despite some shortcomings in some of the elements of the pipeline, which shows the feasibility of a predominantly symbolic, rather than purely statistic, approach to content extraction, especially in the case of a morphologically complex language such as Arabic.

Second, a partial syntactic analysis reveals itself to be sufficient to extract a reasonable amount of information from corpus texts. This is encouraging, since it is expected that a fuller match between syntax and semantics (especially when a fuller argument extraction component is developed, which includes nominalization, a highly prominent feature in Arabic texts) can bring significant improvements.

Third, the results of the project demonstrate that a very complex pipeline of several independent projects can work provided a consistent way to link chains can be found. This stresses the importance of developing links between existing lexical resources in order to capitalize on their interconnection.

Further developments in the project will include — besides optimization, in order to allow researcher for tests on larger data sets, and refinements in the evaluation stage, to allow a finer assessment of the contribution of the single components— strategies for anaphora resolution, analysis of inter-clausal relations (in order to avoid wrong interpretations of counterfactuals and other "possible worlds" structures) and the development of links to other existing resources, such as the Arabic version of VerbNet.

REFERENCES

[1] http://www.ldc.upenn.edu/Catalog/docs/LDC2010T13/atb1-v4.1-taglist-conversion-to-PennPOS-forrelease.lisp

[2] T. Buckwalter, Buckwalter Arabic Morphological Analyzer Version 1.0. Philadelphia:Linguistic Data Consortium, 2002.

[3] T. Buckwalter and D. Parkinson, A Frequency Dictionary of Arabic. London and New York: Routledge, 2011.

[4] S. Green and J. DeNero, "A class-based agreement model for generating accurately inflected translations", in ACL '12, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1, pp. 146-155.

[5] C. Fellbaum, "WordNet and wordnets", in: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 2005, pp. 665-670

[6] S. Green and Ch. D. Manning, "Better Arabic parsing: baselines, evaluations, and analysis", in COLING '10, Proceedings of the 23rd International Conference on Computational Linguistics, pp. 394-402.

[7] M. Boella, F. R. Romani, A. Al-Raies, C. Solimando, and G. Lancioni, "The SALAH Project: segmentation and linguistic analysis of hadīth arabic texts. information retrieval technology lecture notes" in Computer Science vol. 7097, Springer, Heidelberg, 2011, pp 538-549.

[8] Ch. M. Meyer and I. Gurevych, "What psycholinguists know about chemistry: aligning Wiktionary and WordNet for increased domain coverage", in Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 2011, pp. 883-892.

[9] E. Niemann and I. Gurevych, "The people's web meets linguistic knowledge: automatic sense alignment of wikipedia and wordnet", in: Proceedings of the International Conference on Computational Semantics (IWCS), Oxford, United Kingdom, 2011, pp. 205-214.

[10] E. Wolf and I. Gurevych, "Aligning Sense Inventories in Wikipedia and WordNet", in: Proceedings of the First Workshop on Automated Knowledge Base Construction (AKBC), Grenoble, France, 2010, pp. 24-28.

[11] H. A. Salmoné, An Advanced Learner's Arabic-English Dictionary. Beirut: Librairie du Liban, 1889.

[12] N. Habash, O. Rambow, and R. Roth, "MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization", in Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, 2009, pp. 102-109.

[13] I. Turki Khemakhem, S. Jamoussi and A. Ben Hamadou, "Integrating morpho-syntactic features in English-Arabic statistical machine translation", in Proceedings of the Second Workshop on Hybrid Approaches to Translation", Sofia, Bulgaria, 2013, pp. 74-81.

[14] M. Daoud, D. Daoud and Ch. Boitet, "Collaborative construction of Arabic lexical resources", in Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, 2009, pp. 119-124.

[15] S. Izwaini, "Problems of Arabic Machine Translation: evaluation of three systems", in Proceedings of the International Conference on the Challenge of Arabic for NLP/MT, The British Computer Society (BSC), London, 2006, pp. 118-148.

[16] Arabic Treebank Guidelines, http://www.ircs.upenn.edu/arabic/guidelines.html. Accessed October 2013.

[17] https://github.com/anastaw/Meedan-Memory. Accessed October 2013.

 ISBN: 978-1-61208-311-7