مجلة القُلزم للدراسات التطبيقية

مجلة القلزم للدراسات التطبيقية
QUZUM JOURNAL FOR APPLIED STUDIES

ISSN: 1858 - 9553

علمية دولية مُحكمة - تصدر بالشراكة مع جامعة دنقلا - السودان

في هذا العدد:

RSBCRSC

العدد السابع - ذو الحجة 1445هـ - يونيو 2024م

فهرسة المكتبة الوطنية السودانية - السودان

مجلة القُلزم للدراسات التطبيقية

Qulzum Journal for Applied Studies

الخرطوم : مركز بحوث ودراسات دول حوض البحر الأحمر 2024

تصدر عن دار آريثيريا للنشر والتوزيع - السوق العربي الخرطوم - السودان

ردمك: 1858-9553

ترسل الأوراق العلمية على العنوان التالي
هاتف: +249121566207 - +249910785855
بريد إلكتروني: rsbcrsc@gmail.com
السودان - الخرطوم - السوق العربي
عمارة جي تاون - الطابق الثالث

موجهات النشر

تعريف المجلة:

مجلة (القُلزم) للدراسات التطبيقية، مجلة علمية مُحَكمة تصدر عن مركز بحوث ودراسات دول حوض البحر الأحمر - السودان بالشراكة مع جامعة دنقا- السودان. تهتم المجلة بالبحوث والدراسات العلمية والمواضيع ذات الصلة بدول حوض البحر الأحمر.

موجهات المجلة:

1. يجب أن يتسم البحث بالجودة والأصالة، وألا يكون قد سبق نشره قبل ذلك.
2. على الباحث أن يقدم بحثه من نسختين. وأن يكون بخط (Traditional Arabic) بحجم 14 على أن تكون الجداول مرقمة وفي نهاية البحث وقبل المراجع على أن يشار إلى رقم الجدول بين قوسين دائريين ().
3. يجب ترقيم جميع الصفحات تسلسلياً وبالأرقام العربية بما في ذلك الجداول والأشكال التي تلحق بالبحث.
4. المصادر والمراجع الحديثة يستخدم أسم المؤلف، اسم الكتاب، رقم الطبعة، مكان الطبع، تاريخ الطبع، رقم الصفحة.
5. المصادر الأجنبية يستخدم اسم العائلة (Hill, R).
6. يجب ألا يزيد البحث عن 30 صفحة، وبالإمكان كتابته باللغة العربية أو الإنجليزية.
7. يجب أن يكون هناك مستخلص لكل بحث باللغتين العربية والإنجليزية على ألا يزيد على 200 كلمة بالنسبة للغة الإنجليزية. أما بالنسبة للغة العربية فيجب أن يكون المستخلص وافياً للبحث بما في ذلك طريقة البحث والنتائج والاستنتاجات، مما يساعد القارئ العربي على استيعاب موضوع البحث وبما لا يزيد عن 300 كلمة.
8. لا تلزم هيئة تحرير المجلة بإعادة الأوراق التي لم يتم قبولها للنشر.
9. على الباحث إرفاق عنوانه كاملاً مع الورقة المقدمة (الاسم رباعي، مكان العمل، الهاتف، البريد الإلكتروني).

نأمل قراءة شروط النشر قبل الشروع في إعداد الورقة العلمية.

بسم الله الرحمن الرحيم

الحمد لله رب العالمين، والصلاة والسلام على سيدنا محمد وعلى آله وصحبه أجمعين

وبعد:

القارئ الكريم ،،،

السلام عليكم ورحمة الله وبركاته.. نطل على حضراتكم من نافذة جديدة من نوافذ النشر العلمي، وهي **مجلة القُلزم للدراسات التطبيقية**، ونحن في غاية السعادة والمجلة تصل عددها السابع بفضل الله تعالى ومنته.

القارئ الكريم:

هذه المجلة تصدر بالشراكة مع جامعة دنقلا، وهي إحدى الجامعات السودانية الفتية التي وضعت بصمات مميزة في مسيرة البحث العلمي، وهذا العدد هو السابع في إطار هذه الشراكة العلمية التي تأتي في إطار استراتيجية مركز بحوث ودراسات دول حوض البحر الأحمر، في تفعيل الحراك العلمي والبحثي داخل السوداني وخارجه.

القارئ الكريم:

هذا العدد يشتمل على العديد من البحوث والدراسات المهمة ذات البعد النظري والتطبيقي، ولضمان نجاح واستمرارية هذه المجلة بإذن الله تعالى، نأمل أن يرفدنا الباحثون بمزيد من إسهاماتهم العلمية المميزة.

مع خالص الشكر والتقدير للجميع

أسرة التحرير

المحتويات

Lexical Semantical Analysis Methods & Techniques For Academic Arabic Text

■ A. Mosab Ibrahim Alrasheed Seed Ahmed
■ Dr. Eltyeb Elsamani Abdelgabar Elsamani

Abstract:

This paper provides an inductive study of the challenges and methods for extracting lexical and semantical information from Arabic text. Arabic is a complex and diverse language that poses unique challenges for natural language processing tasks, particularly in the area of lexical semantics extraction. The main challenges include the lack of resources and tools available for the language, as well as the morphological complexity of the Arabic language. To overcome these challenges, researchers have developed a number of methods, including the use of morphological analyzers, corpus-basd methods, distributional semantics, and deep learning techniques. A simplified algorithm that describes the steps for lexical semantical analysis for academic Arabic has been created, the paper argues that these methods have shown promising results and have the potential to improve the performance of natural language processing tasks for Arabic language.

Keywords: Arabic NLP, morphological complexity,corpus-based methods, Sentiment Analysis, Word Sense Disambiguation.

طرق وتقنيات التحليل الدلالي اللغوي لنصوص اللغة العربية الاكاديمية

أ. مصعب إبراهيم الرشيد سيد أحمد
طالب دكتوراه - كلية الدراسات العليا - جامعة النيلين.

د. الطيب السماني عبدالجبار
أستاذ مشارك - كلية علوم الحاسوب وتقانة المعلومات - جامعة النيلين

المستخلص:

تقدم هذه الورقة دراسة استقراءيه عن أساليب وطرق وتحديات استخراج المعلومات الدلالية من النص العربي. اللغة العربية تعتبر لغة معقدة ومتنوعة وتشكل تحديات فريدة لمهام معالجة اللغات الطبيعية، وخصوصا في مجال استخراج المعلومات اللغوية والدلالية، وتشمل التحديات الرئيسية نقص الموارد والأدوات المتاحة للباحثين، فضلا عن التعقيد الصياغي للغة.

للتغلب على هذه التحديات، قام الباحثون بتحليل عدد من الطرق، بما في ذلك استخدام المحللات الصرفية والطرق القائمة على أنظمة المعاجم اللغوية، وتلك التي تعتمد على انظمه الدلالات التوزيعية والاحصائية وتقنيات تعلم الآلة ومقارنتها من حيث دقه النتائج. يرى الباحثون في هذه الورقة أن هذه الطرق قد أظهرت نتائج مؤكدة ولديها الإمكانية لتحسين أداء مهام معالجة اللغة الطبيعية للعربية. كماتم تطوير خوارزمية مبسطه تصف خطوات التحليل الدلالي العامة للنصوص العربية. وأخيرا خلصت الورقة لأن هذه الأساليب أظهرت نتائج واعدة ولديها القدرة على تحسين أداء مهام معالجة اللغات الطبيعية للغة العربية.

الكلمات المفتاحية: المعالجة اللغات الطبيعية للغة العربية، التعقيد الصرفي، تحديد أجزاء الكلامTop of FormBottom of Form, تحليل العاطفة، توضيح المعاني.

1. Introduction:

Arabic language is a complex, rich and diverse language, which makes it challenging to extract lexical semantical information from it. Despite the challenges, extracting lexical information from Arabic text has become

an important field of research in natural language processing, and with the emergence of Large Language Models (LLMs) this need is becoming increasingly more important. In this paper, the challenges of extracting lexical semantical information from Arabic text and the methods used to overcome these challenges will be discussed, finally an algorithm is proposed to streamline these tasks based on open source standards and solutions available.

2. Challenge:

One of the main challenges in extracting lexical semantical information from Arabic text is the lack of resources and tools available for the language. Unlike other languages such as English, Arabic has a much smaller amount of text data and resources available for analysis. This makes it difficult to train machine learning models for Arabic text and limits the types of techniques that can be used for lexical semantics extraction.

Another challenge is the morphological complexity of the Arabic language. Arabic is an inflected language, which means that the form of a word changes to indicate grammatical functions such as tense, gender, and numerical descriptions (كَتَبَ كَتَبَتْ، تَكْتُبَانِ). This complexity makes it difficult to identify the root form of a word and extract its meaning[1].

3. Lexical semantic extraction methods:

A common approach for extracting lexical-semantic information from Arabic text is morphological analyzers, which can help identify the root form of a word and extract its meaning. Researchers have also used techniques such as corpus-based methods and distributional semantics, which leverage large amounts of text data to extract lexical information.[3]

Another promising approach is using deep learning techniques to extract lexical- information from Arabic text. techniques such as neural network based models and transformer architectures, both have been used with promising results. These models have been found to be particularly effective in capturing the complex and diverse nature of Arabic language. [3]

In addition to the aforementioned methods, there are also other methods that are used to overcome the challenges of lexical-semantic extraction, For example, some researchers use WordNet-like resources and Word2Vec models that were trained on Arabic corpus and could extract lexical information at with persistant results. [6]

In conclusion, extracting lexical-semantic information from Arabic text is challenging However, researchers have developed a number of methods to overcome these challenges, including the use of morphological analyzers, corpus-based methods, distributional semantics, and deep learning techniques. These methods have shown promising results and have the potential to improve the performance of natural language processing tasks for Arabic.

4. Techniques for Arabic lexical semantic extraction:

There are several techniques that have been developed for extracting lexical semantic information from Arabic text. Some of these include:

i. Morphological analysis: This technique involves analyzing the structure of words in order to identify the root form of a word and extract its meaning. Researchers have developed a number of morphological analyzers for Arabic that can help overcome the challenges of the morphological complexity of the language[2].

   Example: "يكتبون" - meaning "they write" in English.

   a) Tokenization: The word is tokenized to isolate it for analysis.

   b) Identification of the Root: The analyzer identifies the trilateral root "ك-ت-ب" (k-t-b), which pertains to writing or inscribing.

   c) Pattern Recognition: The word "يكتبـون" follows a pattern that can be mapped to a specific morphological pattern. For Arabic, the pattern here "يفعلـون" (yaf'aluun) is recognized, which is a common pattern for the plural third-person masculine present tense.

d) Extraction of Grammatical Features:

- Tense: Present tense
- Number: Plural
- Gender: Masculine
- Person: Third person
- Mood: Indicative (used for factual statements and positive beliefs)
- Voice: Active (the subject of the sentence is performing the action)

e) Lemmatization: The word is reduced to its lemma form "كتب" (kataba), which means "to write."

f) Semantic Information: Additional semantic information may be provided, such as synonyms, antonyms, and usage examples, depending on the analyzer's capabilities.

i. Corpus-based methods: These methods use large amounts of text data to extract lexical-semantic information. They can include techniques such as co-occurrence analysis, collocation analysis, and distributional semantics which leverage large amounts of text data to extract lexical-semantic information[4].

ii. Deep learning techniques: Researchers have used neural network-based models and transformer architectures to extract lexical-semantic information from Arabic text. These models have been found to be particularly effective in capturing the complex and diverse nature of the Arabic language. [4]

iii. WordNet-based methods: Some researchers use WordNet-like resources to extract lexical-semantic information from Arabic text, by mapping words and concepts to a structured lexical database.

iv. Word embeddings models: Techniques such as word2vec are used to train models on Arabic corpus, allowing extraction of lexical-semantic information from the text. This teqnique is widely used in transformer based LLMs particuallry the GPT familty of models.

v. Hybrid methods: Hybrid methods can be used to extract lexical-semantic information, it combines several techniques to leverage the advantages of each method.

vi. Semantic Role Labeling: a technique that can be used to extract the meaning of a sentence by identifying the roles of words and entities in a sentence.

vii. Named Entity Recognition: a technique used to extract named entities and their attributes from text.

These techniques have been proven to be effective in extracting lexical-semantic information from Arabic text, with each one having its own advantages and limitations. The choice of technique depends on the specific task, resources, and data availability, as well as the desired level of accuracy.

5. Lexical Semantics Analysis:

A lexical semantic network is a type of computational model that represents the relationships between words and concepts in a language. These networks are typically built using a set of techniques and algorithms that allow them to learn about the structure of a language by analyzing large amounts of text data for that particual language. The basic building block of a lexical semantic network is a node, which represents a token, word or a concept. These nodes are connected to one another by edges, which represent the relationships between the words or concepts. For example, in a lexical semantics network of English, the word “cat” might be connected to the word “animal” by an edge representing the concept of “is-a” (i.e., “cat is-a animal”).

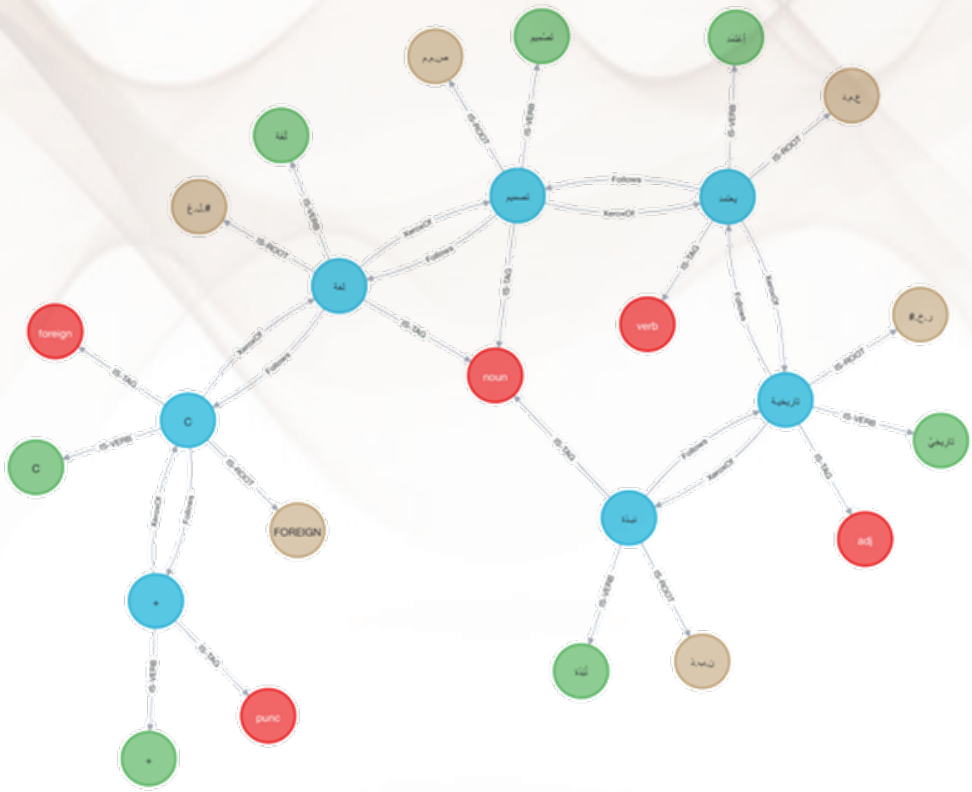

Fig 1 - Graph representation of a lexical semantical network of an Arabic sentence

There are several different ways to create a lexical semantics network, but a common approach is to use techniques from natural language processing (NLP) to extract information from text data and build the network automatically. Here are some steps that can be used to create a lexical semantic network:

I. Collect a large corpus of text data: The first step in creating a lexical semantic network is to collect a large corpus of text data that will be used to learn about the structure of a language. This data can be collected from a variety of sources, such as books, articles, and websites.

II. Pre-processing: The collected data should go through the text pre-processing step in order to clean the text data and make it suitable for the analysis. This might include steps such as tokenization, stemming, and removing stop-words.

III. Extracting relationships: Next, relationships between words are extracted from the text data. This can be done using a variety of NLP techniques, such as dependency parsing or co-occurrence analysis. These techniques can be used to identify patterns in the data that indicate relationships between words, such as synonymy, antonymy, or hypernymy-hyponymy[2].

IV. Creating the network: Once relationships between words have been extracted, they can be used to create the network. The extracted relationships can be represented as edges between nodes in the network, with each node representing a word or concept.

V. Evaluating and refining the network: The final step is to evaluate the network and refine it as needed. This can be done by comparing the network to a gold standard set of relationships and adjusting the network as needed. Additionally, the network can be fine-tuned by adding more data and repeating the extraction and creation steps.

Lexical semantic networks can be used for a variety of natural languages processing tasks such as word sense disambiguation, text classification, and machine translation. The process of creating lexical semantic networks can be complex and computationally expensive, but it allows to have a rich representation of the language that can be used in various NLP applications.

The table below provides a comparission across NLP tasks for the most popular frameworks used in Arabic langugage research.

Feature	UralicNLP	Qutuf	AlKhalil Morpho Sys 2	MADA +TOKAN
Open-Source	Yes	Yes	No	Yes
Functionality	Lemmatization, Morphological Analysis, Generation	Morphological Analysis, POS Tagging	Morphological Analysis, Syntactic Analysis	Tokenization, Diacritization, Disambiguation, POS Tagging, Stemming, Lemmatization
Morphological Coverage	Good (MSA), Limited Dialects	Comprehensive (MSA)	Excellent (MSA, Explores Dialects)	Comprehensive (MSA)
Disambiguation	Not directly supported	Limited Capabilities	N/A	Advanced Disambiguation
Stemming & Lemmatization	Both Stemming & Lemmatization	N/A	N/A	Both Stemming & Lemmatization
Syntactic Analysis	N/A	N/A	Provides Syntactic Analysis	N/A
Accuracy & Reliability	Good (MSA), Requires Evaluation for Dialects	Well-established Reputation	Recognized for High Accuracy	Highly Regarded (Accurate Analysis with Disambiguation)
Performance	Efficient (Basic Tasks), Might Need Optimization	Performance Not Optimal for complex tasks	N/A (Commercial)	Optimized for Efficient Processing
Customization	Open-source (Code Modification)	Limited Customization Options	Limited Customization Options (Commercial)	Highly Customizable (Open-source)
Documentation & Support	Good Documentation & Community Support	Limited Documentation Available	Limited Documentation (Commercial, Might Require Paid Support)	Excellent Documentation & Active Community Support

Table 1- popular NLP frameworks and libararies comparision

6. A Lexical analysis algorithm:

A simplified algorithm for performing lexical analysis on Arabic text that relies on existing NLP tools including the Stanford NLP toolset using python goes as follows:

1. Input: A string of Arabic text (sentence).
2. Output: A list of Arabic words with their root form, POS tags, and lexical information such as lemma, singular, gender, and meaning.

```
from morphology.database import MorphologyDB
from morphology.analyzer import Analyzer

# First, we need to load a morphological database.
# Here, we load the default database which is used for analyzing
# Modern Standard Arabic.
db = MorphologyDB.builtin_db()

analyzer = Analyzer(db)

analyses = analyzer.analyze('برمجة')

for analysis in analyses:
    print(analysis, '\n')
```

3. Initialize an empty list to store the analyzed words
4. Normalize the input text by removing diacritics, punctuation, and non-Arabic characters.

```
from utils.dediac import dediac_ar

sentence = "الحَاسُوب هُو جِهَاز يَقُوم بِإِدْخَال الْبَيَانَات ، مُعَالَجَتِهَا وَإِخْرَاج النَّتَائِج ."
print(sentence)

sent_dediac = dediac_ar(sentence)
print(sent_dediac)
```

Input.الحَاسُوب هُو جِهَاز يَقُوم بِإِدْخَال الْبَيَانَات، مُعَالَجَتِهَا وَإِخْرَاج النَّتَائِج :

Output.الحاسوب هو جهاز يقوم بإدخال البيانات، معالجتها وإخراج النتائج :

5. Tokenize the input text into a list of words by using a tokenizer that can handle Arabic script, such as an ICU-based tokenizer

```
from tokenizers.word import simple_word_tokenize

sentence = "هَلْ ذَهَبْتَ إِلَى المَكْتَبَةِ؟"
print(sentence)

sent_split = simple_word_tokenize(sentence)
print(sent_split)
```

: Input.الحَاسُوب هُو جِهَاز يَقُوم بِإِدْخَال الْبَيَانَات، مُعَالَجَتِهَا وَإِخْرَاج النَّتَائِج

: Output,[‹الحَاسُـوب›, ‹هُـو›, ‹جِهَـاز›, ‹يَقُـوم›, ‹بِإِدْخَـال›, ‹الْبَيَانَـات›, ‹،›, ‹مُعَالَجَتِهَـا› ‹وَإِخْـرَاج›, ‹النَّتَائِـج›, ‹.›]

6. Run the tokenized text through a morphological analyzer to identify the root form of the words, POS tags, and other morphological information.

Lex	POS Tags (Part-of-Speech Tags)	Diacritized	Word
ما	pron_rel	ما	ما
هِيَ	pron	هِيَ	هي
لُغَة	noun	لُغات	لغات
بَرْمَجَة	noun	البَرْمَجَة	البرمجة

Table 2 – arabic sentenc analysis for diacritized and part of speech generation

7. For each word in the tokenized text:

   A. Use a morphological analyzer to extract its root form, lemma, and morphological information such as gender, number, and POS tags

   B. Use a lexical resource, such as a WordNet-like resource or a large corpus, to extract lexical information such as synonyms, antonyms, and word senses.

C. Append the word, root form, POS tags, lemma, morphological information, and lexical information to the list of analyzed words.

Gloss	Transliteration	Word
what	REL_PRON/ما	ما
it;they;she	PRON_3FS/هِيَ	هي
[.language+[fem.pl	NOUN/لُغ NSUFF_FEM_PL/ات+	لغات
the+programming [.+[fem.sg	DET/ال+بَرْمَج/ NOUN+ةَ NSUFF_FEM_SG/	البرمجة

Table 3 – arabic sentenc glossary and buckwalter transliteration

8. Return the list of analyzed words

9. End

This algorithm is a simplified version and there can be variations to it based on the context and requirements of the specific task. Morphological analysis can be a complex task as it involves several subtasks such as diacritization, segmentation, POS tagging, and more. Some morphological analyzers can handle multiple subtasks at once, which can simplify the process of lexical analysis. Additionally, WordNet-like resources are not always available for Arabic language, so researchers can opt for other resources such as thesauruses, dictionaries, or even their own hand-crafted resources.

7. Available Resources & tools:

There are several resources available for researchers and practitioners interested in extracting lexical-semantic information from Arabic text[7]. Some of these include:

I. Corpora: A number of Arabic corporas have been developed for natural language processing research, including the Arab Corpus, the Egyptian

Arabic Dialect Corpus, and the Modern Standard Arabic Corpus. These corpora can be used to train machine learning models for lexical-semantic extraction.

II. Morphological analyzers: There are a multiple morphological analyzers available for Arabic, such as the MADAMIRA[8] tool and the Alkhalil[9] morphological analyzer, which can be used to analyze the structure of Arabic words and identify the root form.

III. WordNet-like resources: Some lexical resources for Arabic, such as AWN (Arabic Word Net), can be used to extract lexical-semantic information from Arabic text.

IV. Pre-trained models: BERT-Arabic and similar models can be used for Arabic NLP tasks such as lexical-semantic extraction

V. Evaluation datasets: Researchers can use evaluation datasets to evaluate the performance of their models, for example, the Arabic Lexical Simplification Corpus, the Arabic Dependency Treebank, and the Arabic Named Entity Recognition Corpus.

VI. Research papers and journals: There are several research papers and journals that focus on natural language processing and lexical-semantic extraction in Arabic, such as the Journal of King Saud University - Computer and Information Sciences and the Journal of Intelligent Systems.

VII. Conferences and workshops: There are also several conferences and workshops that focus on natural language processing and lexical-semantic extraction in Arabic, such as the Conference on Natural Language Processing and Computational Linguistics and the Workshop on Arabic Corpora and Processing Tools.

VIII. Online communities: Communities such as the Arabic NLP community, which provide a platform for researchers and practitioners

to share information, ask questions, and collaborate on projects related to Arabic natural language processing.

These resources can be used to help researchers and practitioners learn about the latest methods and techniques for extracting lexical-semantic information from Arabic text, as well as to access data and tools that can be used for their own research and projects.

8. Conclusion:

Arabic lexical semantics analysis is an important field of research in natural language processing. The aim of lexical semantic analysis is to extract rich information about words and concepts from a text, such as the root form, lemma,POS tags, morphological information, and lexical information. The Arabic language presents unique challenges for lexical semantic analysis due to its complexity, such as morphological richness, script and diacritics, and lack of resources.

This paper aimed to provides an overview of the challenges and methods used to overcome these challenges for Arabic lexical-semantic analysis. It discussed how the morphological complexity of Arabic can be addressed by using morphological analyzers which can identify the root form of a word and extract its meaning. Additionally, it presented how corpus-based methods and deep learning techniques can be used to extract lexical semantic information from Arabic text, in addition to word embedding models and WordNet-based methods. The paper also highlighted the importance of using evaluation datasets and the need to use large corpora for training.

To conclude, lexical-semantic analysis of Arabic is a challenging task but researchers have developed a number of methods to overcome the challenges. These methods have shown promising results and have the potential to improve the performance of natural language processing tasks for Arabic.

9. References:

(1) Computational Linguistics, Speech and Image Processing for Arabic Language - Challenges in Arabic Natural Language Processing (pp.59-83) 2018

(2) Natural Language Processing with Python by Steven Bird, Ewan Klein, and Edward Loper -.

(3) Arabic Computational Morphology: Knowledge-based and Empirical Methods by Dipti Misra Sharma -

(4) Arabic Natural Language Processing by Michael A. Diab and Kareem Darwish

(5) Arabic Natural Language Processing (arabicnlp.org) –

(6) The Arabic WordNet (arabicwordnet.org)

(7) Arabic NLP Community (https://groups.google.com/g/arabic-nlp)

(8) MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic By Arfath Pasha, Mohamed Al-Badrashiny & others

(9) AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer By Mohamed Boudchiche, Azzeddine Mazroui

دار آريثيريا للنشر والتوزيع
Arrythria for Publishing and Distribuation

ISSN: 1858-9553 ردمك