

Arabic Natural Language Understanding as a Sustainable AI Research

Mustafa Jarrar
Birzeit University, Palestine

Artificial Intelligence

Opportunity
to invest in
Arabic
as industry
and
sustainable
R&D

The 4th
Revolution

Opportunity
for
developing
countries to
compete and
impact

A close-up photograph of a woman with long, dark, wavy hair. She is looking upwards and slightly to the left with her eyes closed and her mouth open, as if she is singing or yawning. The background is dark and out of focus. The entire image is enclosed within a white, decorative scalloped border.

A black and white photograph of a person's arm resting on a dark surface. The arm is bent, and a leather cuff bracelet is visible on the wrist. In the background, there is large, bold Arabic text.

Huge Market



400
Million
users
(as a first
language)

1.6
Billion
users
(liturgical/
second
Language)

Other
users for
economic
and security
...

Consequences of not supporting Arabic in ICT

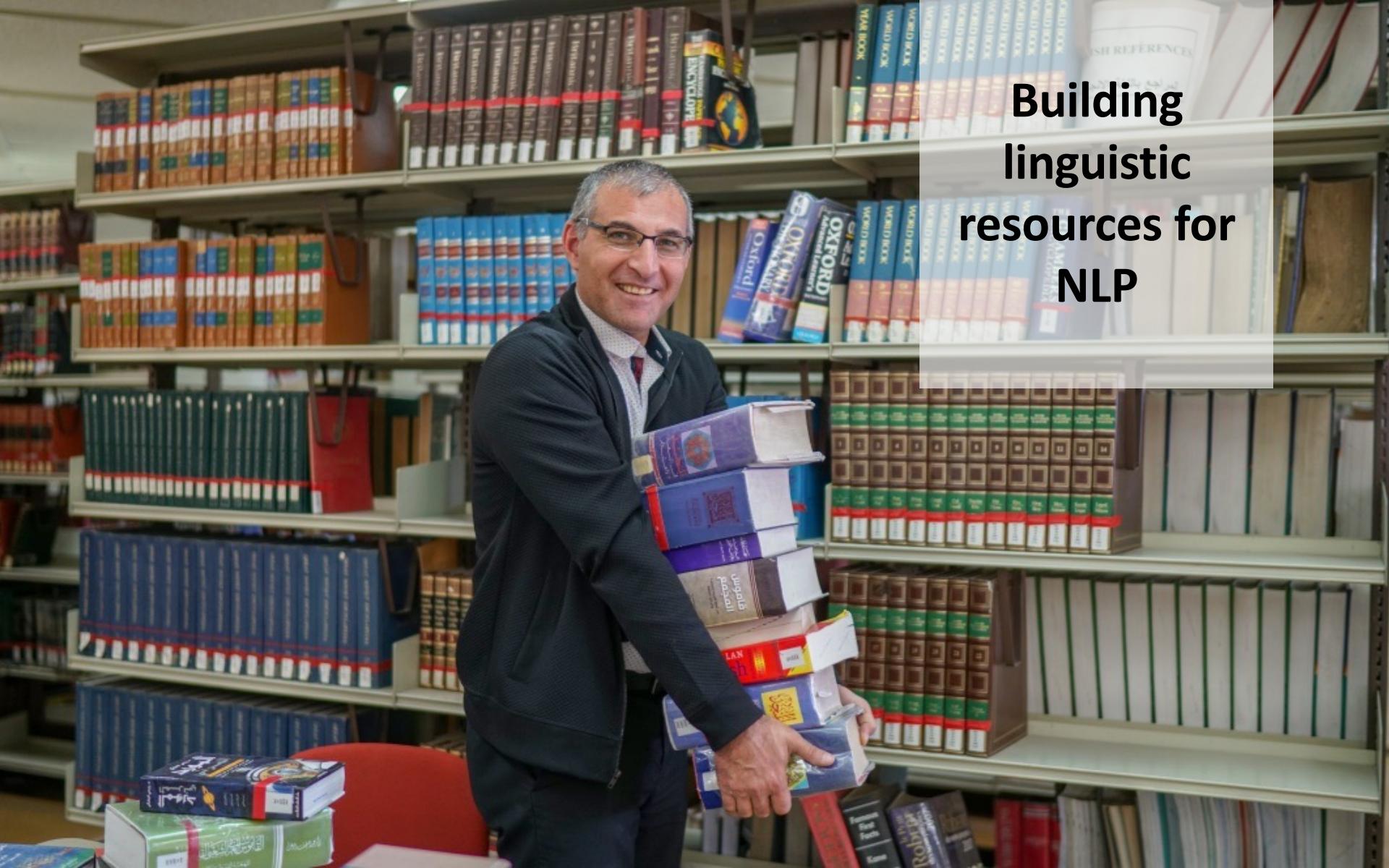
**On Education, culture, economy, politics, people with special
needs,**

Why Lexical Resources!

- ❖ The importance of lexical resources (dictionaries, thesauri, wordnets, linguistic ontologies) is increasing in many application areas, such as:
 - NLP tasks and applications
 - Information search and retrieval
 - Multilingual big data
 - Multilingual semantic web
 - Data integration
 - among many others.
- ❖ Lack of Arabic Lexical resources for human use!
- ❖ Lack of Arabic Lexical resources for NLP!

Digitize, Collect, Build, then clean and link

- Make available online for people
- Make available through APIs for NLP applications

A photograph of a man with grey hair and glasses, wearing a dark jacket over a patterned shirt, standing in a library. He is holding a large, tall stack of books against his chest with both hands. The books are of various sizes and colors, including blue, purple, red, and white. The background consists of several rows of bookshelves filled with books.

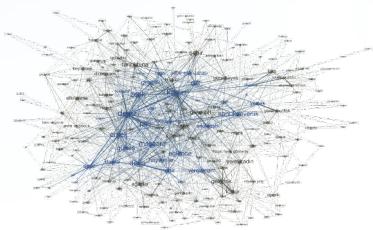
Building linguistic resources for NLP

ISH REFERENCES



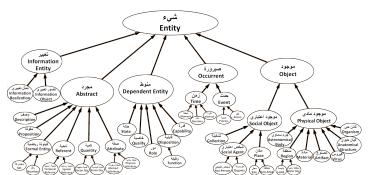
Lexical Resources at Birzeit University

Lexicographic Database



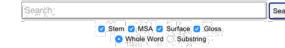
150 lexicons
Very large Arabic-multilingual database

Arabic Ontology



Formal Arabic Wordnet
with ontologically clean content

Dialect Corpora



Annotated corpora
each word is annotated
with many morph features



Big Linguistic Data Graph

<https://ontology.birzeit.edu>

Lexicographic Search Engine

The Lexicographic Database

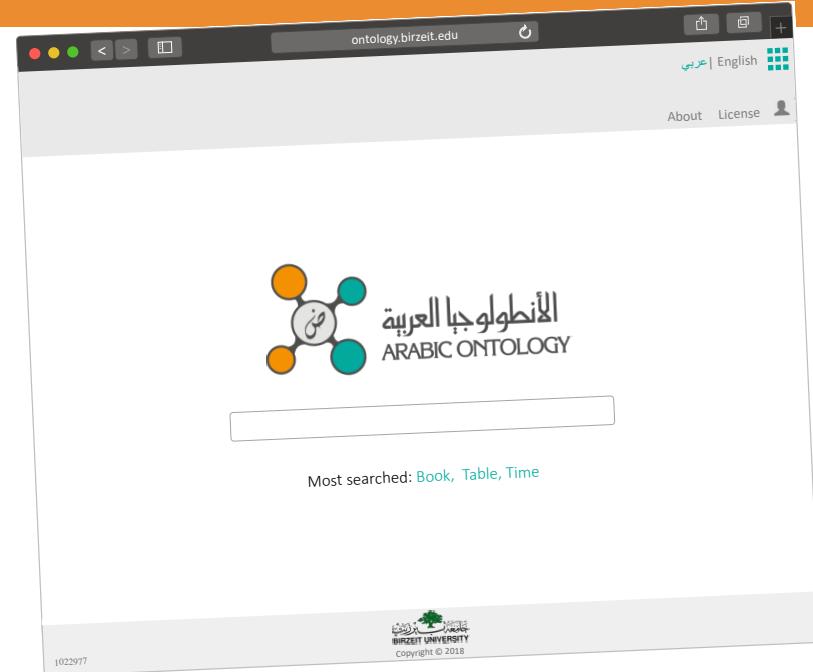
- The largest lexicographic Arabic database
- Contains most lexicon types: glossaries, thesauri, bi/trilingual dictionaries, morph datasets, **Arabic Ontology**, and more.
- Covers most domains: science, technology, law, business, art, philosophy, ...

The screenshot shows a web browser window for ontology.birzeit.edu. The search bar contains the word "attribute". Below the search bar are three checkboxes: "Translations", "Synonyms", and "Definitions", all of which are checked. The main content area displays search results for "Predicate | Attribute".
The first result is "السمول عند المطهفي هو المكتوب به في القضية الجملية دون الشرطية، اما في الشرطية فيسمى ثالثاً ففي قوله: زيد كريم، زيد هو الم موضوع، وكريم هو المسمول، والموضوع والمسمول عند المطهفي [المزيد](#)".
The second result is "الصلة في المثل الدال على بعض الحالات، او الصلة التي تكون عليها الشيء: كالسراويل، والبياض، والعبء، والجبل، الخ... و الصلة عند التحريف هي النعت، امام المقابل، واسم المفهول، والصلة [المزيد](#)".
The third result is "خاصة ينبع بها شخص او شيء بواسطة لفظ محدد في الجملة" with a link to "Philosophy Lexicon (V1.8.2) ©".
The fourth result is "صلة تتعذر عن قيمة خاصية مجزأة لنفسها، ما يكرر في صفة مجزأة لخاصية المعلمة عن الإنسان: كرم" with a link to "Arabic WordNet ©".
The fifth result is "صلة تتعذر عن قيمة خاصية مجزأة لنفسها، ما يكرر في صفة مجزأة لخاصية المعلمة عن الإنسان: كرم" with a link to "Arabic WordNet ©".
The bottom of the page includes navigation links for "Propositions attributes", "attribute predicate", "attribute", "flavor", "difference", and "character". It also features a footer with the university's logo and copyright information: "Academy of the Arabic Language Cairo", "Al Amira Printing House", "Academ Portal", "Browse Online", and "Copyright © 2018".

<https://ontology.birzeit.edu>

Lexicographic Search Engine

- **Free access to people:** students, translators, researchers, Arabic learners ...
- **API accessible** for NLP applications.
- **Done over 9 years (No funding!)**



<https://ontology.birzeit.edu>

Reference:

Mustafa Jarrar, Hamzeh Amayreh: **An Arabic-Multilingual Database with a Lexicographic Search Engine.** NLDB 2019. Pages(234--246), LNCS 11608, Springer. 2019.

Lexicographic Search Engine

- **Search 150 lexicons** for definitions, synonyms, specialized translations, morphology, ontology [3,4] ...
- **Accurate!** compared with machine translation.
- **The first of its kind!** e.g., there are no similar search engines for English lexicons!



Some Statistics

Currently!

Category	Lexical Concepts	Lexical entries	Synsets	Translations pairs	Glosses	Semantic relations
Total (Millions)	1.1 M	2.4 M	1.8 M	1.5 M	0.7 M	0.5 M
Sub Counts	1,100K Arabic 1,100K English 200K French 3K Others 1,300K Single-word 1,000K Multi-word	800K Arabic 800K English 200K French 50K Others	1,000K English-Arabic 300K English-French 200K French-Arabic	400K Arabic 300K English 1K Others	170K Sub-super links 29K Part-of links 260K Has-Domain links 30K Other links	

For more, see [3]

Accessible

To download and access NLP data,
corpora, tools and services

RESTful web services

Ask us for an API Key!

NLP Library

Download and access NLP data, corpora, tools and services

We developed hundreds of RESTful web services (APIs) for other third-party software developers to directly download and access our lexicographic databases, corpora, and tools. Some are listed on this page. Please get in touch with us for details.

The "sampleKey" API token allows you to use all web services but with a limited number of requests (enough to try it!), if you need more requests, you need to [Request API Token](#).

APIs:

- + Search 150 dictionaries (synonyms, translations, glosses)
- + Arabic Ontology
- + Dialect Corpora
- + Something to be announced
- + Something to be announced
- + Something to be announced
- + Retrieve Morphology Information (MSA and Dialects)
- + Arabic Processing Tools
- + Autocomplete Service (Arabic and English)
- + Retrieve data source information

Arabic Ontology

- To enable machine understand semantics.
- Classification of the meanings of the Arabic terms, specified in D. Logic
- Benchmarked to scientific advances rather than to speakers' naïve beliefs as wordnets do.

The screenshot shows a web interface for an Arabic ontology. At the top, there's a navigation bar with tabs for 'Ontology' (which is active), 'Dictionaries', and 'Morphology'. There are also buttons for 'Translations', 'Synonyms', and 'Definitions'. A search bar is at the top right, along with language selection ('En') and user account icons.

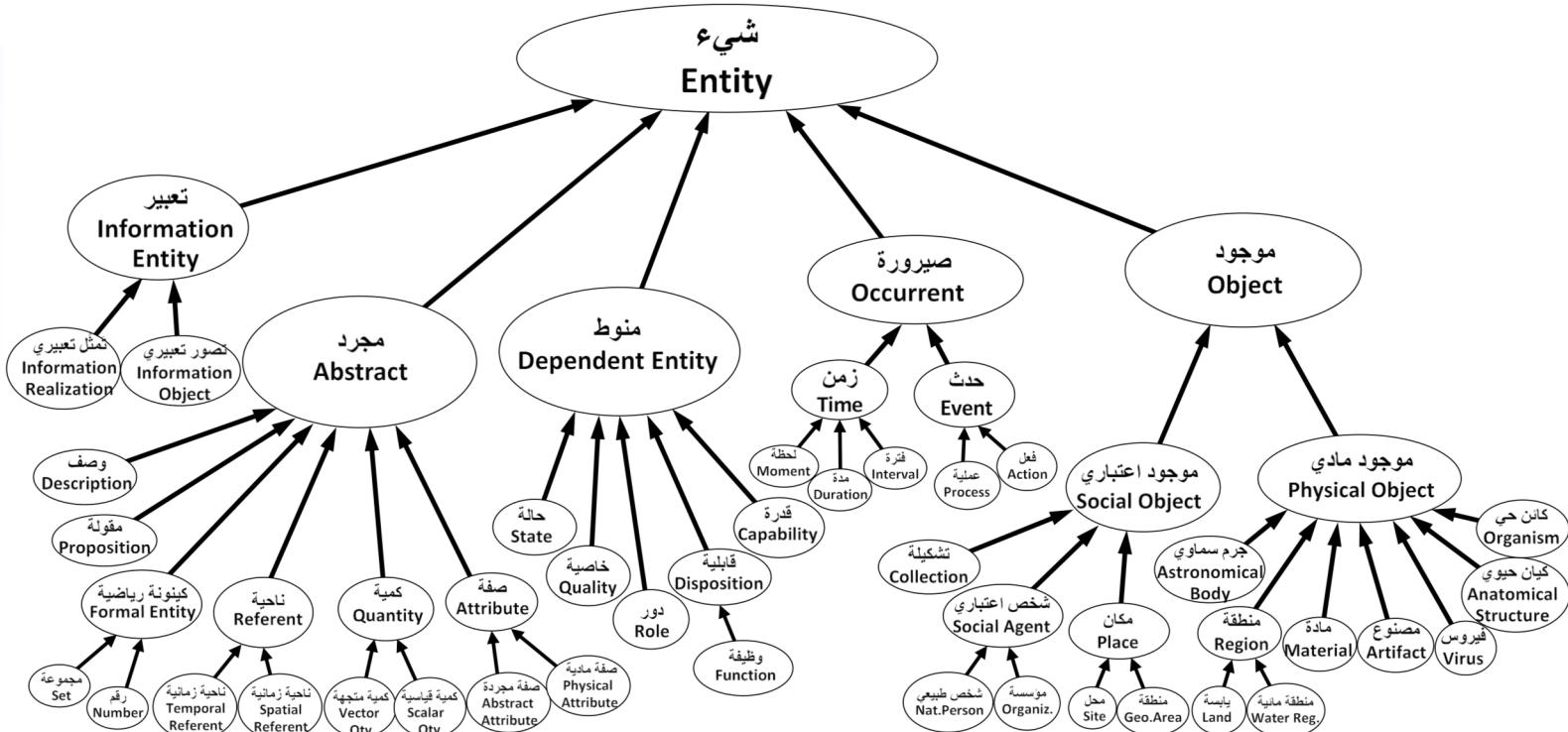
The main content area displays several entity types with their definitions and examples:

- شيء | كائن | Entity**: Whatever existed or will exist, and can be realized or imagined. Example: كل شيء على ما يرام. ID: 293198.
- موجود | كان | قائم | حقيقة | واقعي | شيء | ذات | قوّة**: An entity that is wholly and independently present in time, and is realized either for its concrete or social existence. Example: يختلف إدراكنا لاي موجود لاختلاف ما يميز نوعه من الصفات الجوهرية. ID: 293200. TypeOf: {Entity}
- صيغة | حدث | حادث | وقوع | آخر**: An entity realized by the time of its happening. Example: لا يمكن فهم أي حدث بشكل منفصل عن الإطار الزمني له. ID: 293202. TypeOf: {Entity}
- متصل | معتقد | متسلق | متزوج**: An entity realized by the time of its happening. Example: طول المبني متصل بوجود المبني وإلا فلا طول له. ID: 293201. TypeOf: {Entity}
- مجرد | تجريد | غير مادي | نظري**: An entity exists only in mind, cannot be measured or socially realized, and does not have a location.

At the bottom left, there's a footer with the number 1022977. On the right, there's a logo for Birzeit University and the text "Copyright © 2018".

<https://ontology.birzeit.edu/concept/293198>

Arabic Ontology

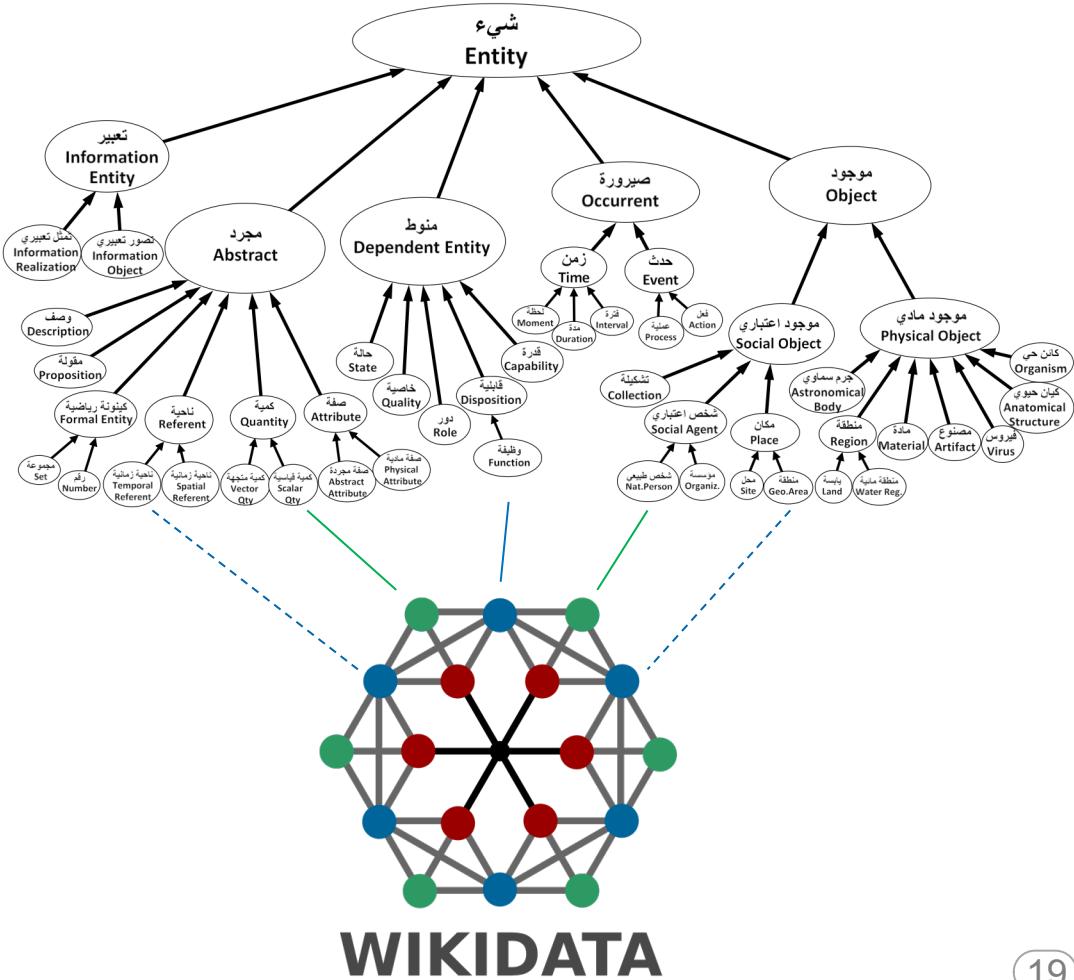


Based on:

Mustafa Jarrar: *The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content*. Applied Ontology Journal, IOS Press. (Forthcoming).

Linking the Ontology with Knowledge Graphs

- Every concept in the ontology is linked with a node in the Wikidata
- In this way, one can access the world knowledge as an Arabic knowledge graph
- So, we can build applications like chatbots and question answering, ...



Dialect Corpus

<http://portal.sina.birzeit.edu/curras>

- To enable machines understand dialects.
- We collected a corpus written in Palestinian dialect (60k words).
- Described and annotate each word with 16 tags.



Based on:

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, Nasser Zalmout: Curras: An Annotated Corpus for the Palestinian Arabic Dialect. Journal Language Resources and Evaluation. Volume(51), Issue(3). Springer. 2017

Dialect Corpus

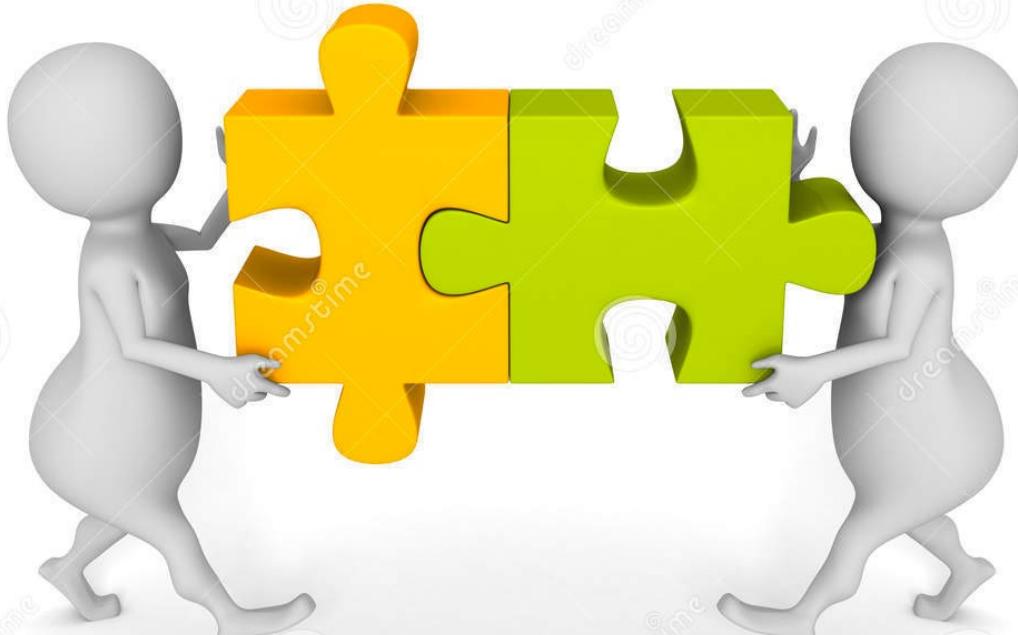
<http://portal.sina.birzeit.edu/curras>

برهان الدين بيرزيت
جامعة بيرزيت
BIRZEIT UNIVERSITY

كلمات بحث

Word Stem MSA Surface Gloss
Whole Word Substring

About												Search												Publications												Download												News												Free Ideas											
بقولكوا	قول	قال	قال	1	قال	1	قال	1	قال	1	قال	بقولكوا	b/PROG_PART+A/IV 1S+qwl/IV+kw/IVSUF F_DO:2P	said;say;be_said; (for_example)_[CALIMA]	Verb	1	M	بنتنا زي كعك العيد بتتفق بسرعة. العريض : والله انتو عيلة ما مشترىكوا بقمعة سجارة ، اخر اشي بقولكوا اياه غروس مالبورو شو قلتوا ؟ الام : اسمع ولا ، خاي ا	زيون يا بقتل امي . مثال : عنا ينطلع بالنكسي . . . ايه ازمة . يتسأل الشوفير شوف س ليش مازمة ؟ بقولكوا ما هو اليوم الخميس هيك . ينطلع السبت يا زلمة شو القص	ير شوف س ليش مازمة ؟ بقولكوا ما هو اليوم الخميس هيك . ينطلع السبت يا زلمة شو القصه ليش مازمة ؟ بقولكوا يوم السبت عادة هيك ، ينطلع الاثنين له له شوه الازر	مه شو القصه ليش مازمة ؟ بقولكوا يوم السبت عادة هيك ، ينطلع الاثنين له له شوه الازرمه شو في ؟ بقولكوا يوم الاثنين شحال دايما هيك . ينطلع الثلاثاء ول لي	الثلاثاء ول ليش كل هازمة . يوم الثلاثاء يختي معروف فس نفس . ينطلع الخميس ليش هلاقفة الأزماء بقولكوا يوم الخميس دايما هيك ، هذا يوم في الأسبوع طاير . بترو	اط@@صتند@@@ ل@@@ اط@@@ مثال . عنا ينطلع بالنكسي . . . ايه ازمة . يتسأل الشوفير شوف س ليش مازمة ؟ بقولكوا ما هو اليوم الخميس هيك . ينطلع السبت يا زلمة شو القص	الثلاثاء ول ليش كل هازمة . يوم الثلاثاء يختي معروف فس نفس . ينطلع الخميس ليش هلاقفة الأزماء بقولكوا يوم الخميس دايما هيك ، هذا يوم في الأسبوع طاير .																																															



Connecting lexical resources

Lexicon Model for Ontologies: Community Report, 10 May 2016



Final Community Group Report 10 May 2016

Editors:

Philipp Cimiano (Cognitive Interaction Technology Excellence Center, Bielefeld University)

John P. McCrae (Insight Centre for Data Analytics, National University of Ireland, Galway)

Paul Buitelaar (Insight Centre for Data Analytics, National University of Ireland, Galway)

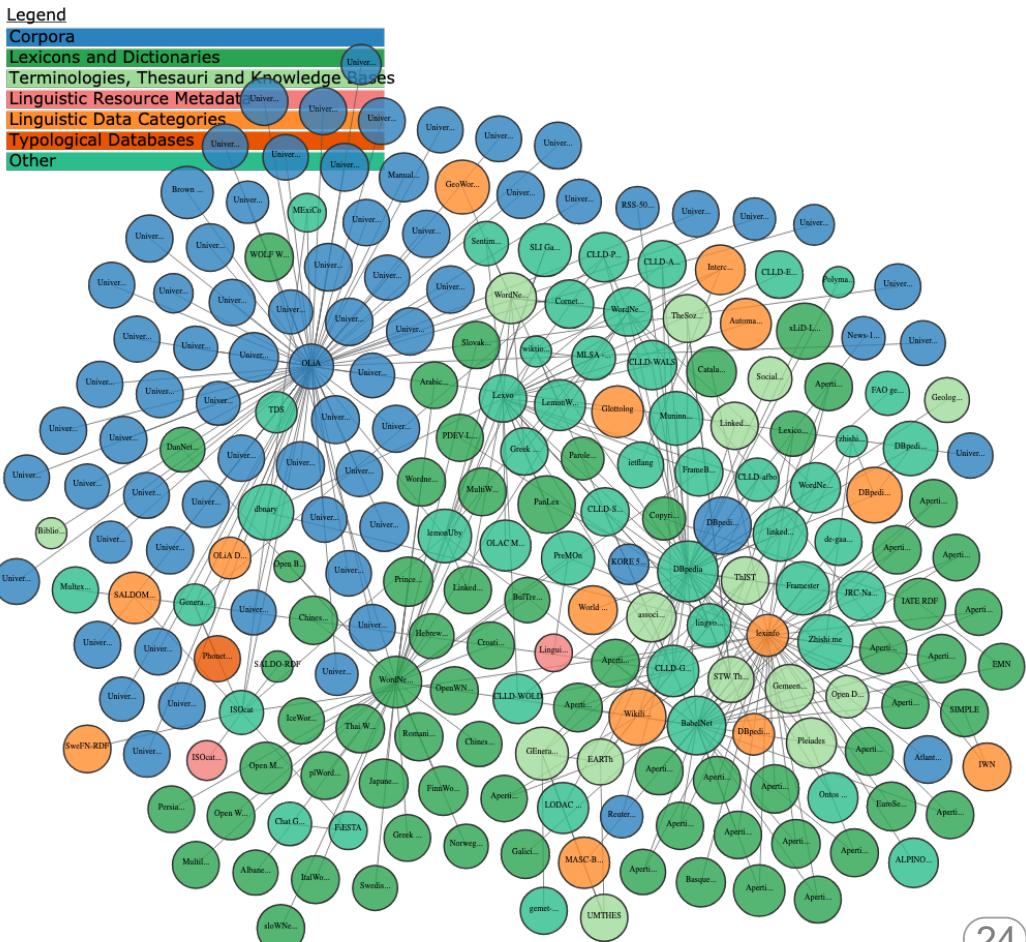
Copyright © 2016 the Contributors to the Lexicon Model for Ontologies: Community Report, 10 May 2016 Specification, published by the Ontology-Lexicon Community Group under the W3C Community Final Specification Agreement (FSA). A human-readable summary is available.

Abstract

This document describes the lexicon model for ontologies (*lemon*) as a main outcome of the work of the Ontology Lexicon (Ontolex) community group.

The Linguistic Linked Open Data Cloud

- A collaborative effort to develop a Linked Open Data (sub-)cloud of linguistic resources.
 - Represent (lexical entries, concepts, synsets, and other) using Lemon RDF model, then interlinked.



Linking All linguistics levels

خمس دكاي بوصل الکمر

CODA: القمر
 POS_ar: ال/أداة تعريف + قمر/اسم
 POS_bw: Al/DET + qmr/NOUN
 Gender: مذكر
 Number: مفرد
 Lemma: قمر 1

المدخلة: قمر 1
 (المصدر: قاعدة بيانات سما)
 اللغة: فصحي حديثة
 نوع المدخلة: اسم
 تصريفات أخرى: قمر، قمر، قمر، قمر، قمر،
 المعنى:
 (AO_51587) moon ◊

Astrophysical Body | جسم فلكي
 A Physical object exists naturally in space
 موجود مادي يتواجد بشكل طبيعي في الفضاء
 النجم هو جرم سمائي متهرج ومشتعل ومضيء بذلك
 example: 293258 TypeOf : (physical object)
 جرم سمائي يدور حول كوكب آخر، لا يشمل ذلك الكواكب التي تدور حول سُمُوس.
 نيل ارمنستونغ هو أول رجل خطى على قمر الأرض
 example: 51587 TypeOf : (Astronomical Body)

CODA: دقائق
 POS_ar: ادقائق/اسم
 POS_bw: dqAyq/NOUN
 Gender: مؤنث
 Number: جمع
 Lemma: دقيقة 1

المدخلة: دقيقة 1
 (المصدر: قاعدة بيانات سما)
 اللغة: فصحي حديثة
 نوع المدخلة: اسم
 صيغة المترافق المؤنث: دقيقة
 صيغة المترافق المونث: دقيقةين / دقيقةين / دقيقة /
 تصريفات أخرى: دقيقة، دقائق، دقائق، دقائق،
 المعنى:
 (AO_293582) minute ◊

Sentence in MSA/Dialect

Morphology Level

Lexical/
Ontology
Semantics



Example Case Study

Word Sense Disambiguation

Moustafa Al-Hajj, Mustafa Jarrar: [ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD](#). In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). PP 40--48, 2021

The problem

The Word Sense Disambiguation (WSD) Task

Given a word in a context, which sense (i.e. meaning) this word denotes?

قصيدة من عيون الشعر

Set of senses

1. عُضو الإبصار في الإنسان والحيوان: له عينان كعَيْنَيُ الصقر - ألا إنما العينان للقلب رائد ...
2. جاسوس، "كان عيناً لدولة أجنبية . بِتُّ العيون : تجسس، راقب - فلان عين على فلان : ناظر عليه
3. أجود كل شيء وأحسن ونفيسه: عيون الفن.
4. حارس: فلان عين على المكان.
5. الحاضر من كل شيء أصبح أثراً بعد عين ...
6. عَيْنُ الماء:- ينبع منه، تُحْلِق الطيور فوق عيون الماء
7. عَيْنُ الشَّيْء:- نفسه، ذاته (تستعمل للتوكيد): جاء القوم أعينهم - كُنَّا في المكان عينه.
8. عَيْنُ العقل:- قدرة ذهنية موروثة على التخييل وتذكر الأحداث.
- 9

WSD has been a challenging task for many years but has gained recent attention due to the advances in contextualized word embedding models such as BERT.

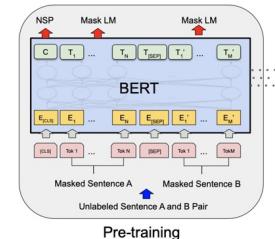
❖ Arabic context-gloss pairs Dataset (167k)

- Extracted from Birzeit University's Lexicographic database
- Annotated target words in context;

Gloss	Context	Label
[CLS] قصيدة من عيون الشعر [SEP] أجود كل شيء وأحسنها ونفيسيه [SEP]		True
[CLS] قصيدة من عيون الشعر [SEP] عين الشيء : نفسه ، ذاته (تستعمل للتوكيد) [SEP]		False
[CLS] جاء القوم أعينهم [SEP] عين الشيء : نفسه ، ذاته (تستعمل للتوكيد) [SEP]		True
[CLS] جاء القوم أعينهم [SEP] أجود كل شيء وأحسنها ونفيسيه [SEP]		False

❖ Three Fine-tuned BERT Models

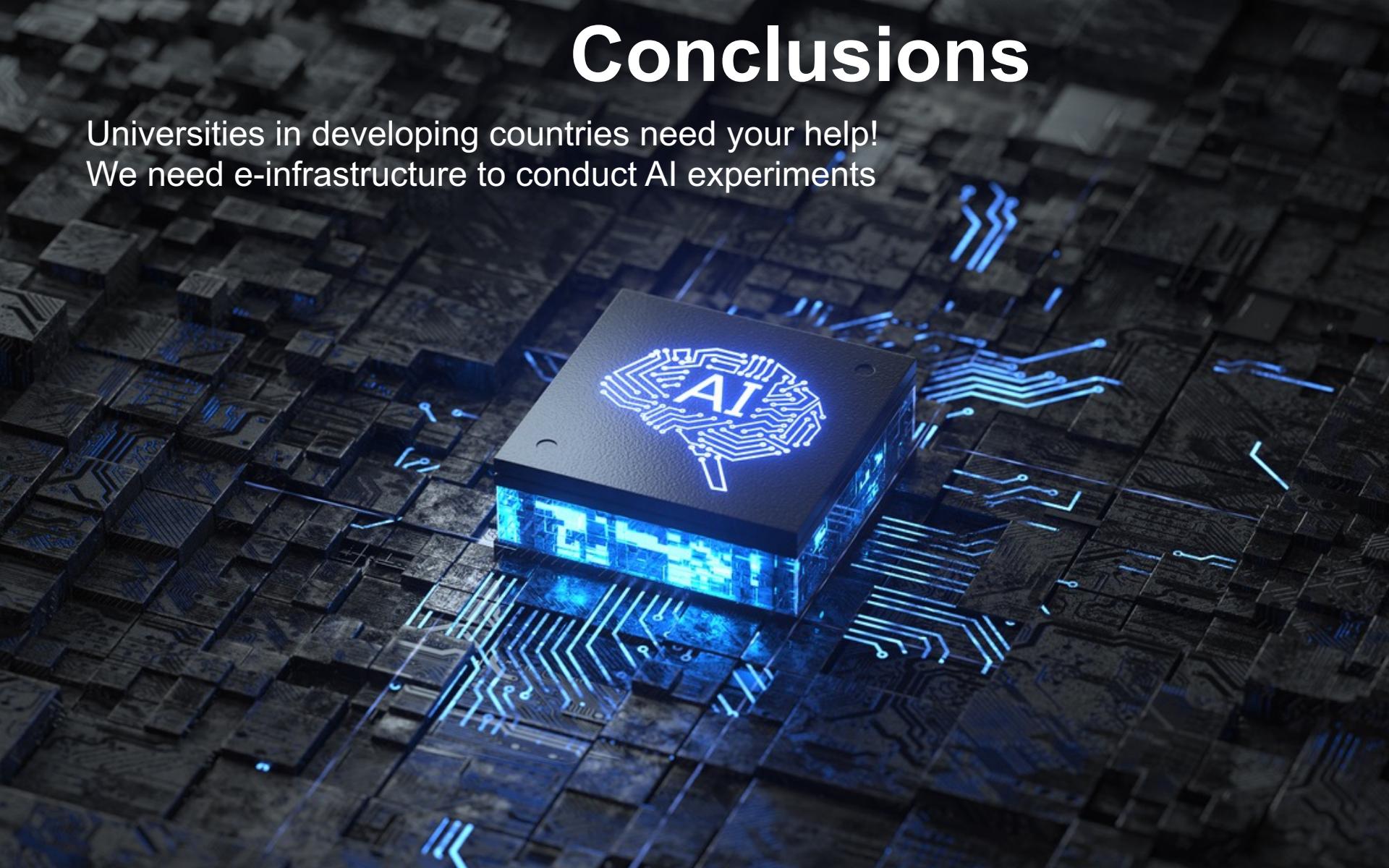
- WSD into **binary sequence-pair classification task**
- **Accuracy 84%**
- 4 types of signals to emphasize target words in context



Conclusions

Universities in developing countries need your help!

We need e-infrastructure to conduct AI experiments



References

1. Moustafa Al-Hajj, Mustafa Jarrar: [ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD](#). In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). PP 40–48, 2021
2. Mustafa Jarrar. [The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content](#). Applied Ontology Journal, IOS Press, 2021.
3. Mustafa Jarrar, Hamzeh Amayreh. [An Arabic-Multilingual Database with a Lexicographic Search Engine](#). Proceedings of the 24th International Conference on Applications of Natural Language to Information Systems (NLDB), 2019.
Mustafa Jarrar: [Search Engine for Arabic Lexicons](#). Proceedings of the 5th Conference on Translation and the Problematics of Cross-cultural Understanding. The Forum for Arab and International Relations. Doha, Qatar. 2018
4. Hamzeh Amayreh, Mohammad Dwaikat, and Mustafa Jarrar. [Lexicon Digitization -A Framework for Structuring, Normalizing and Cleaning Lexical Entries](#). Technical Report, Birzeit University. 2018.
5. Mustafa Jarrar, Fadi Zaraket, Rami Asia, Hamzeh Amayreh: [Diacritic-Based Matching of Arabic Words](#). ACM Asian and Low-Resource Language Information Processing. Volume 18, No 2, Pages(10:1–10:21), ACM, December 2018.
6. Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. [Curras: An Annotated Corpus for the Palestinian Arabic Dialect](#). Journal Language Resources and Evaluation, 51(3):745–775, 2017.
7. Mustafa Jarrar, Nizar Habash, Diyam Akra, Nasser Zalmout: [Building a Corpus for Palestinian Arabic: a Preliminary Study](#). In proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing. Association for Computational Linguistics (ACL), Pages (18-27). October 25, 2014, Doha, Qatar. ISBN: 978-1-937284-96-1
8. Mustafa Jarrar. [Building a Formal Arabic Ontology \(Invited Paper\)](#). In Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. ALECSO, Arab League, 2011

Thank You

Mustafa Jarrar
mjarrar@birzeit.edu