

نحو بناء قواعد البيانات معجمية للفصحى والعامية

مصطفى جرار
جامعة بيرزيت - فلسطين

الذكاء الاصطناعي

فرصة
الاستثمار باللغة
العربية
صناعة وبحث
علمي مستدام

يمثل قلب
الثورة الصناعية الرابعة

فرصة للدول
الصغيرة
للتنافس
والبروز

سوق ضخمة جدا

جهات وأسواق
أخرى
لدواعي اقتصادية
وأمنية

١,٥ مليار
مستخدم للغة
العربية
(كلغة ثانية/دينية)

٣٠٠ مليون
مستخدم في
الوطن العربي
(لغة أولى)

أثر عدم دعم التطبيقات الحديثة للعربية

على التعليم والثقافة والاقتصاد والسياسة وذوي الاجتياحات الخاصة....



❖ ازدادت أهمية المصادر اللغوية (معاجم، مكانز، مسارد، أنطولوجيات..) لبناء التطبيقات الحاسوبية، مثل:

- الترجمة الآلية
- استرجاع البيانات بأكثر من لغة
- فهم وتحليل النصوص المكتوبة والمنطوقة
- التحدث مع الآلة

❖ شح في المصادر اللغوية العربية المحوسبة
تخيلوا أنه لا يوجد قائمة بجميع المدخلات العربية!!!

❖ لا يوجد مصادر لغوية حديثة
مثلاً، لا يوجد مكانز مترادفات تشمل أغلب الكلمات، أو مكانز مصممة بشكل جيد

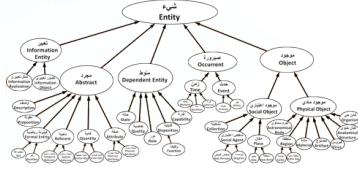
➤ لهذا، يوجد ضعف في دعم التطبيقات للغة العربية

جمع ورقمنة وتنقيح ودمج وتوحيد المصادر اللغوية

- وإتاحتها للجمهور العربي والباحثين ومتعلمي اللغة العربية
- وإتاحتها لمطوري التطبيقات عبر واجهات برمجية (APIs)

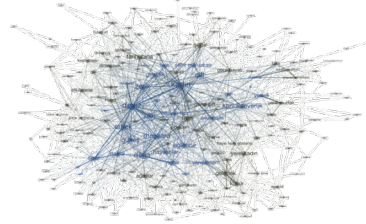
المصادر اللغوية في جامعة بيرزيت

الأنطولوجيا العربية



تصنيف مفاهيم (معاني) الكلمات العربية
Formal Arabic Wordnet
عمل فلسفي حاسوبي لغوي

قاعدة بيانات معجمية



أضخم قاعدة بيانات معجمية
في تاريخ العربية
(تصريفات، اشتقاقات، دلالة)

مدونة للعامية



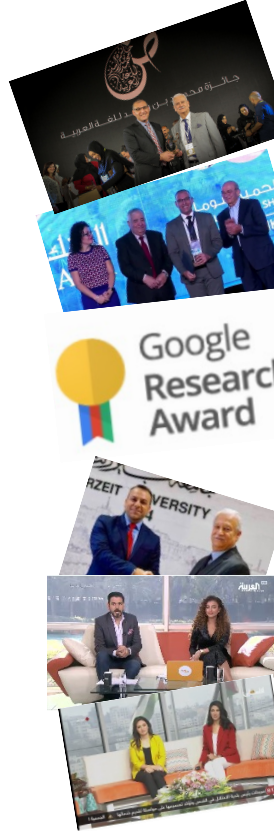
مجموعة ضخمة من النصوص
العامية، تم تصنيف ووسم كل
كلمة فيها بحوالي 16 سمة.

Linguistic Big Data Graph

شبكة بيانات لغوية ضخمة

تجمع المستويات التصريفية والاشتقاقية والدلالية للفصحى والعامية

<https://ontology.birzeit.edu>



قاعدة بيانات معجمية

Lexicographic Database

محرك بحث معجمي

- البحث في 150 معجماً واسترجاع تعريفات، ومترادفات، وفروق لغوية، وترجمات متخصصة.
- دقيق! مقارنة بالمتجمات الآلية.
- الأول عالمياً، فمثلاً، لا يوجد محرك بحث للإنجليزية، يحتوي هذا الكم من المعاجم في قاعدة بيانات واحدة.

ontology.birzeit.edu

محرك بحث معجمي

أكبر قاعدة بيانات لغوية:
الأنطولوجيا العربية + 150 معجم.

150 معجم

معاجم عربية ومتعددة اللغات تم حوسبتها وتوحيدها. حالياً يمكن فقط استرجاع دلالة الكلمات (تعريف، مترادفات، وترجمة) وقريباً سيتم استرجاع التصريفات والاشتقاقات.

الأنطولوجيا العربية

شجرة المفاهيم العربية، تصنيف لمعاني الكلمات وليس الكلمات، مصنفة ومعرفة حسب ما توصلت إليه العلوم، وليس حسب ما شاع بين الناس كما المعاجم. انظر [1].

مفاهيم الأنطولوجيا العليا

المفاهيم الأعلى في شجرة الأنطولوجيا هي المفاهيم الأكثر تجريداً وهي أمهات المعاني الكلمات العربية، تم تقعيدها فلسفياً وتعريفها منطقياً. الشكل التالي يعرض أعلى ثلاثة مستويات في شجرة الأنطولوجيا.

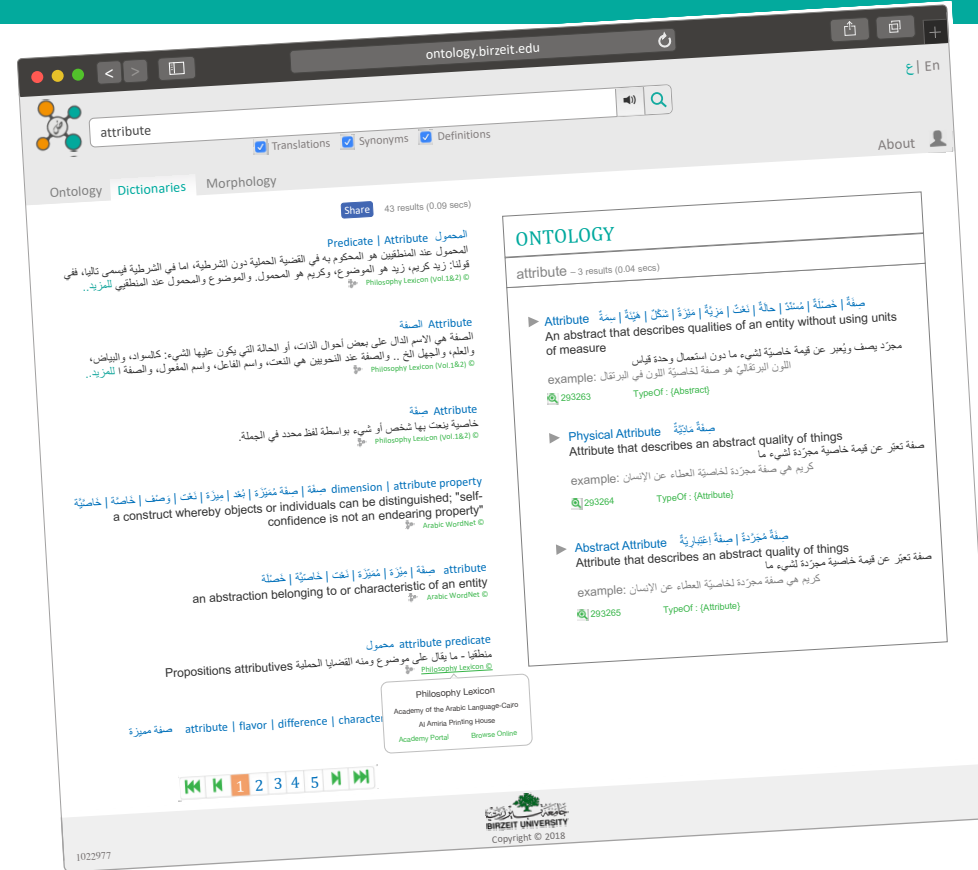
```
graph TD
    Entity[شيء Entity] --> Information[معلومات Information Entity]
    Entity --> Abstract[مجرد Abstract]
    Entity --> Dependent[متوط Dependent Entity]
    Entity --> Occurrent[متصورة Occurrent]
    Entity --> Object[موجود Object]
```

1022977

BIRZEIT UNIVERSITY
Copyright © 2018

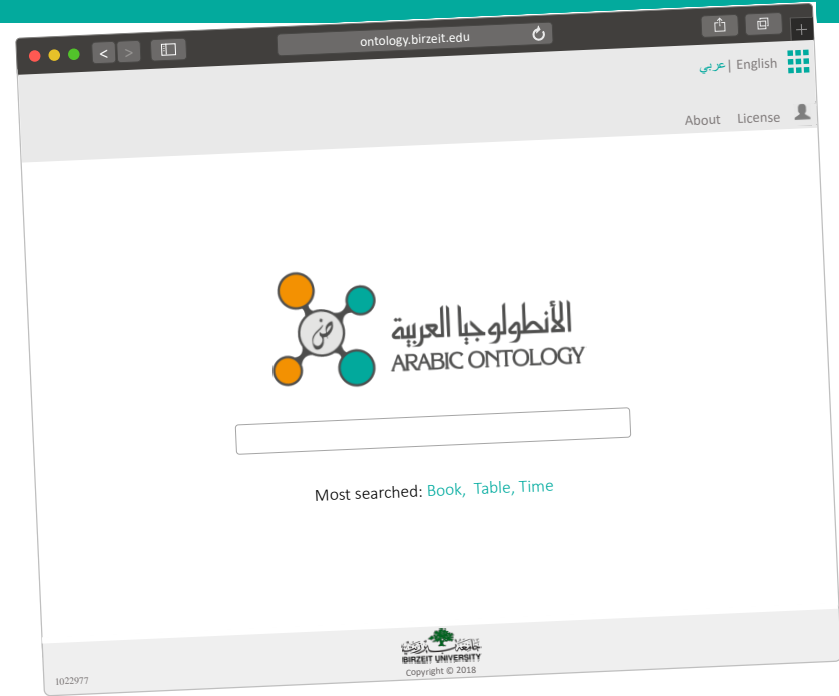
قاعدة بيانات معجمية

- عربية ومتعددة اللغات
- تحتوي على أنواع كثيرة من المصادر اللغوية: معاجم لغوية، ومعاجم ثنائية/ثلاثية اللغة، وقواعد بيانات تصريفية، ومسارد، ومكانز، وفروق لغوية، والأنطولوجيا العربية، وغيرها.
- تغطي معظم المجالات: علوم، وهندسة، وصحة وطب، وفلسفة، وإنسانيات، وفنون، ...



محرك بحث معجمي

- متاحة للعامة: طلبة، مترجمين، باحثين، متعلمي اللغة ...
- متاح لمطوري التطبيقات: عبر واجهات برمجية (APIs)



<https://ontology.birzeit.edu>

Based on:

Mustafa Jarrar, Hamzeh Amayreh: **An Arabic Multilingual Database with a Lexicographic Search Engine**. The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019), Springer. Manchester, UK. 2019

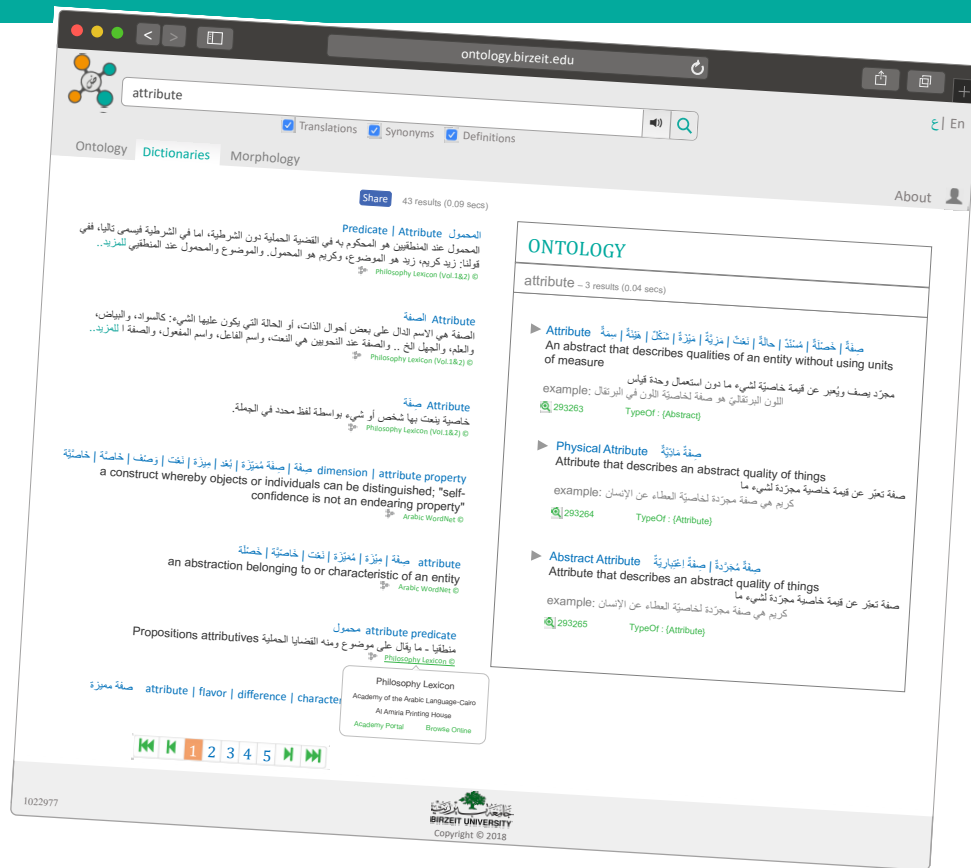
محرك بحث معجمي

تم إنجازه عبر عشرة سنوات

- طباعة يدوية، وإعادة صياغة لبنية المعلومات، وتخزينها في قاعدة بيانات محوسبة.

- عمل طوعي، ولم يتم تمويله من أية جهة داعمة.

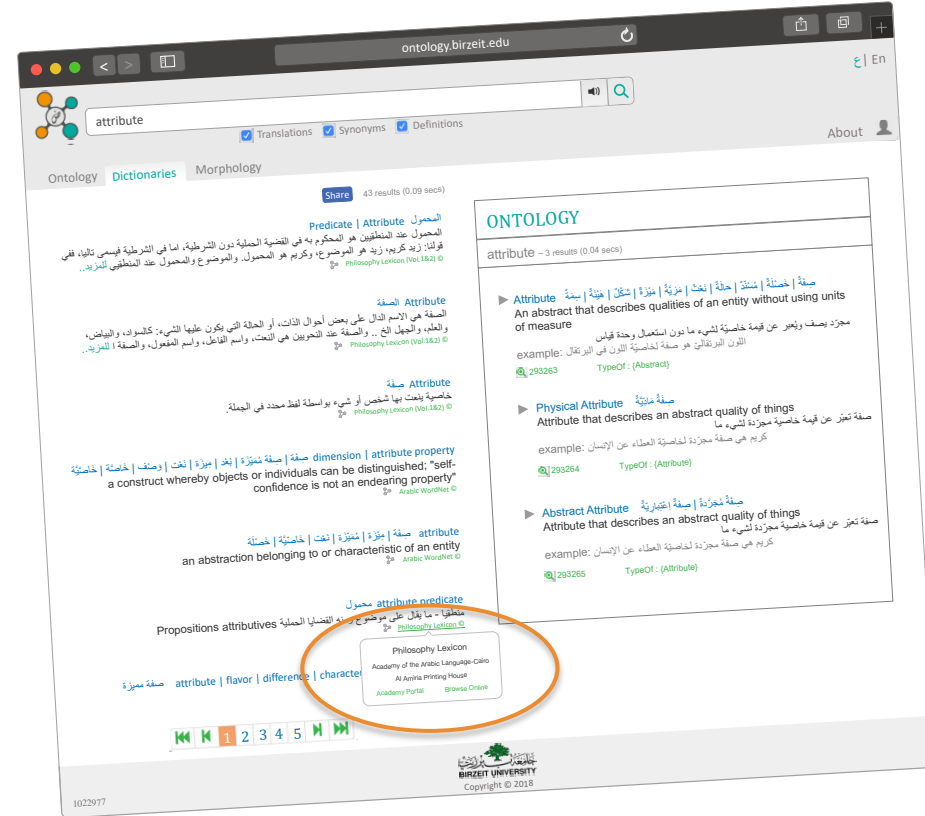
- عمل غير ربحي.



<https://ontology.birzeit.edu>

حقوق الملكية

- تم التواصل مع أصحاب المعاجم وأخذ إذنهم.
- ترويج للمعاجم، وإظهار معلومات المعجم بجانب كل تعريف (المؤلف، الناشر، الصفحة، ...)



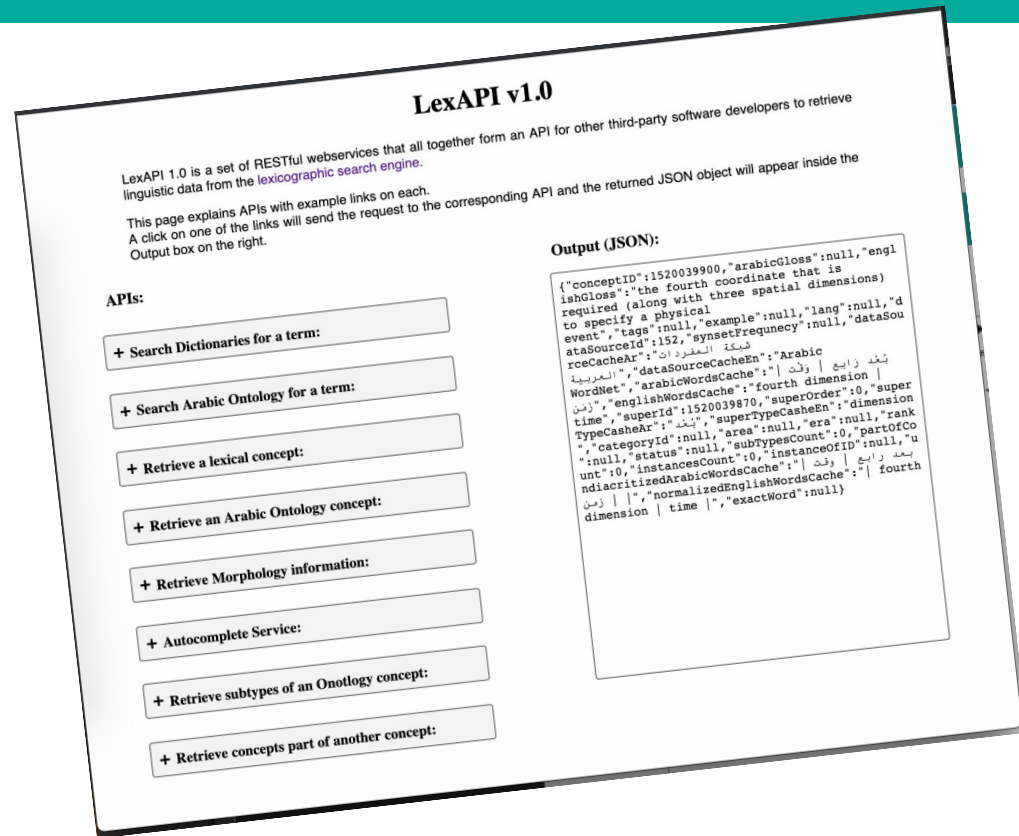
<https://ontology.birzeit.edu>

محرك بحث معجمي

LexAPI

RESTful web services to directly retrieve linguistic data from the lexicographic database (ontology, Morph solutions, synonyms, translations, definitions, etc.).

➤ Request API Token



الأنطولوجيا العربية

Arabic Ontology

(Formal Arabic Wordnet)

الأنطولوجيا العربية

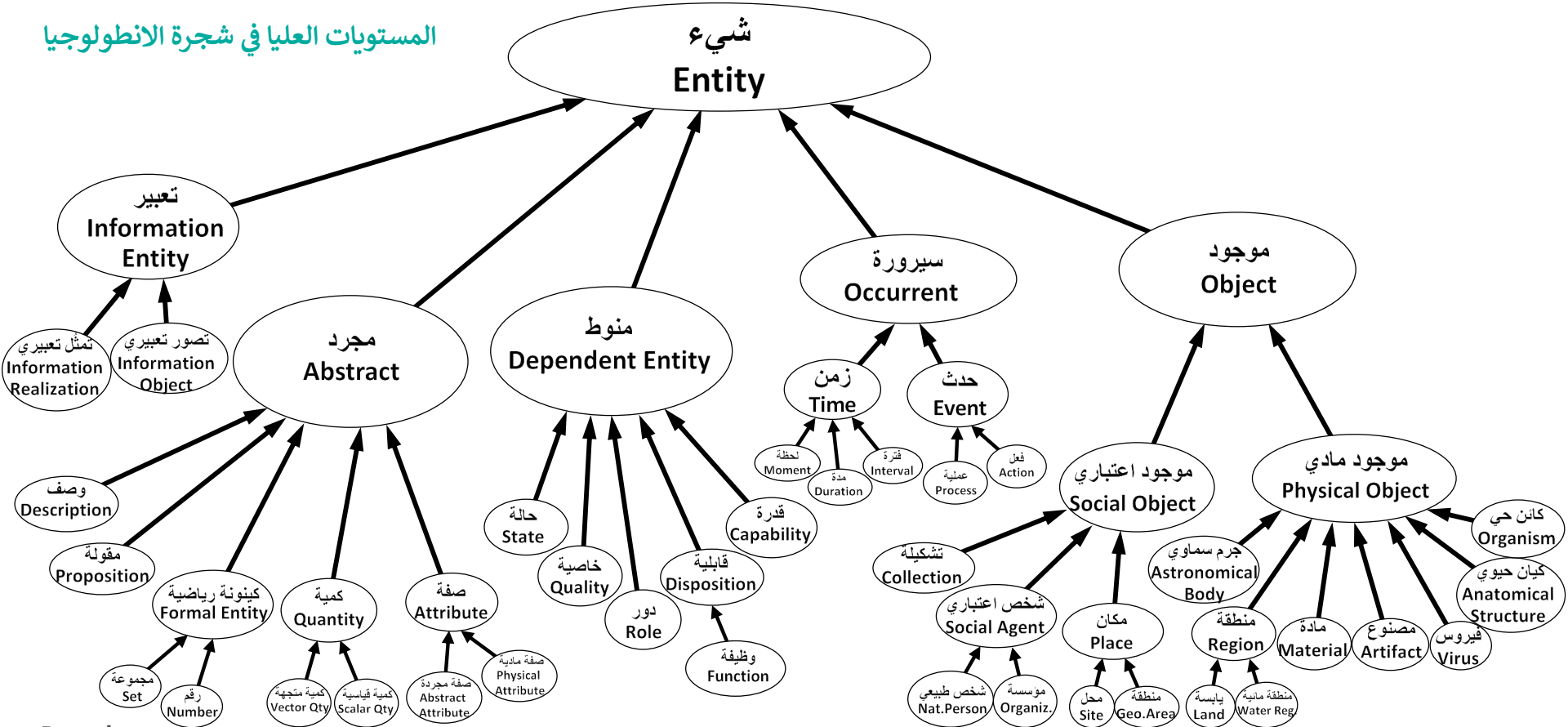
- تستعمل كـ (formal Arabic wordnet)
- عمل حاسوبي وفلسفي ولغوي.
- الأنطولوجيا العربية: وصف معاني الكلمات اللغة العربية، وعلاقات بين هذه المعاني.
- تصنيف مفاهيم (معاني) الكلمات العربية دلاليًا، وتمثيلها بلغة المنطق ليستطيع الحاسوب فهمها.
- تختلف عن المعاجم كون الحكم على صحة المعلومات فيها يعتمد على ما توصلت إليه العلوم، وليس على ما شاع بين الناس.
- مربوطة بشبكة المفردات الإنجليزية (وردنت) والويكي بيانات.

The screenshot shows the Birzeit University Ontology web interface. At the top, there is a search bar and navigation links for 'Translations', 'Synonyms', and 'Definitions'. Below the search bar, the 'Ontology' tab is selected, showing a list of concepts. The first concept is 'Entity' (كائن), with its definition in Arabic and English. The definition in Arabic is: 'شئ | كَيُونَة | كائن Entity: Whatever existed or will exist, and can be realized or imagined أيما وجد أو سيوجد ونستطيع إدراكه أو تخيله'. The definition in English is: 'Object An entity that is wholly and independently present in time, and is realized either for its concrete or social existence شيء له ذات مستقلة بنفسه، وحاضر كلياً في الزمن، ويُدرَك بذاته قياساً أو لذاته اعتباراً'. Below the definition, there is an example: 'example: كلّ شيء على ما يرام 293198'. The interface also shows other concepts like 'Occurrent' (صنوّرة | حدث | حادث | وقع | أمر) and 'Dependent Entity' (مُنَوِّط | مُعْتَمِد | مُتَعَلِّق | مشروط), each with its own definition and example. At the bottom, there is a footer with the Birzeit University logo and copyright information.

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

الأنطولوجيا العربية

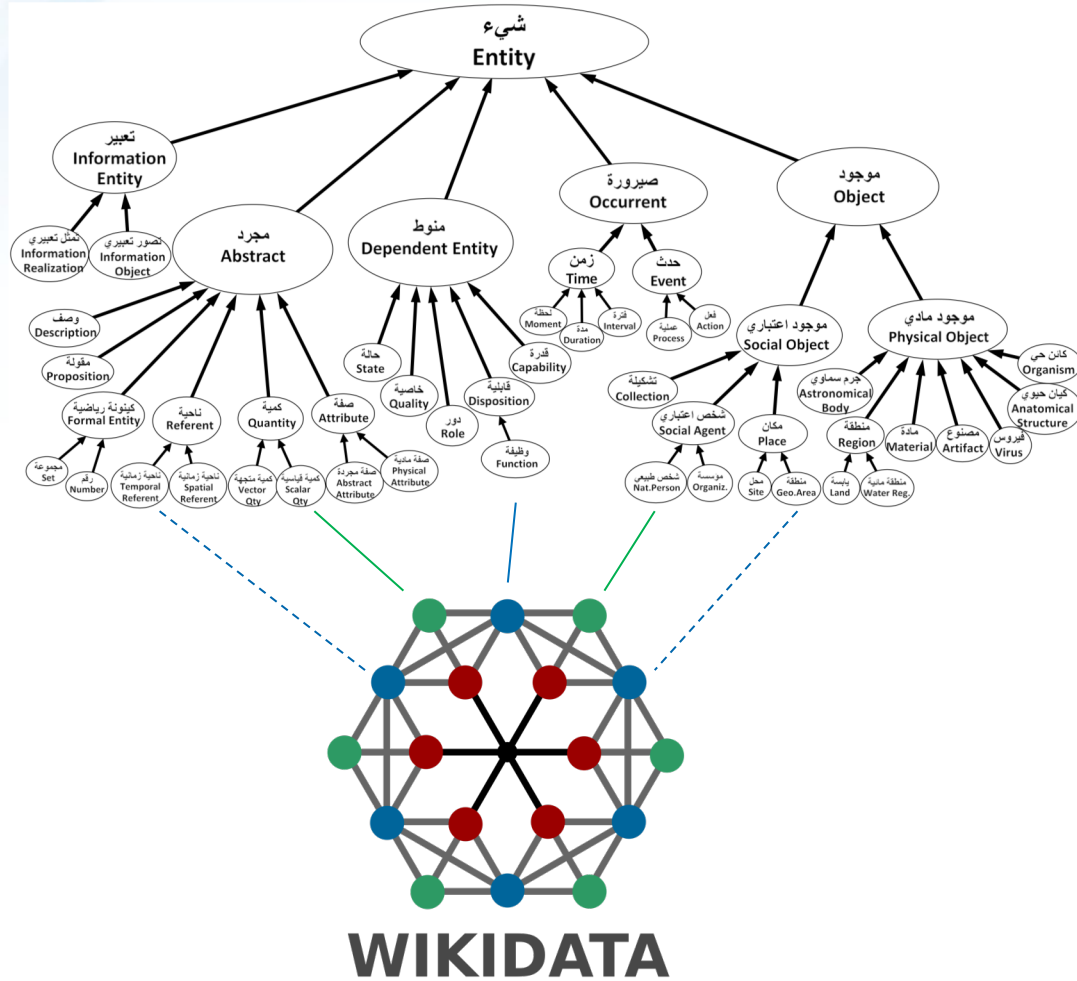
المستويات العليا في شجرة الانطولوجيا



Based on:

Mustafa Jarrar: The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal, IOS Press.2021.

ربط الأنطولوجيا بالويكيبيديا



• تم ربط كل مفهوم بالأنطولوجيا بمقابلة بشبكة بيانات الويكيبيديا (ويكي بيانات).

• وبالتالي، إنشاء شبكة معرفية (Knowledge Graph) مؤصلة بالمفاهيم العربية.

• تطبيقات ذلك مثل، إجابة عن أسئلة معرفية عامة بالعربية.

مدونة اللهجة العامية

Dialectal Corpus

مدونة اللهجة العامية

- عدد ضخم من النصوص المكتوبة باللهجة العامية.
- تصريف كل كلمة بالمدونة ووسمها بحوالي 16 سمة (سوابق، لواحق، جذع الكلمة، فرع الكلمة، المقابل بالفصحى، المقابل بالإنجليزية، ...).
- تحتوي 60 ألف كلمة.



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

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

مدونة اللهجة العامية

تصريف كل كلمة بالمدونة ووسمها بحوالي 16 صفة (سوابق، لواحق، جذع الكلمة، فرع الكلمة، المقابل بالفصحى، المقابل بالإنجليزية، ...).

شو القصة ليش مأزمة؟

Lemma: قصة ١

POS: ال/DET + قص/NOUN + ة/NSUFF_FEM_SG

Person: N/A

Gender: female

Gloss: story;stories_

....

Lemma: شو ١

MSA: ماذا ١

POS: شو/INTERROG_PRON

Person: N/A

Gender: male

Gloss: what;which

....

60k ...

مدونة اللهجة العامية

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

كراس
Curras

Corpus for Palestinian Arabic
مدونة اللهجة العامية الفلسطينية

بقولك

Search

- ☒ Word ☒ Stem ☒ MSA ☒ Surface ☒ Gloss
☐ Whole Word ☒ Substring



About

Search

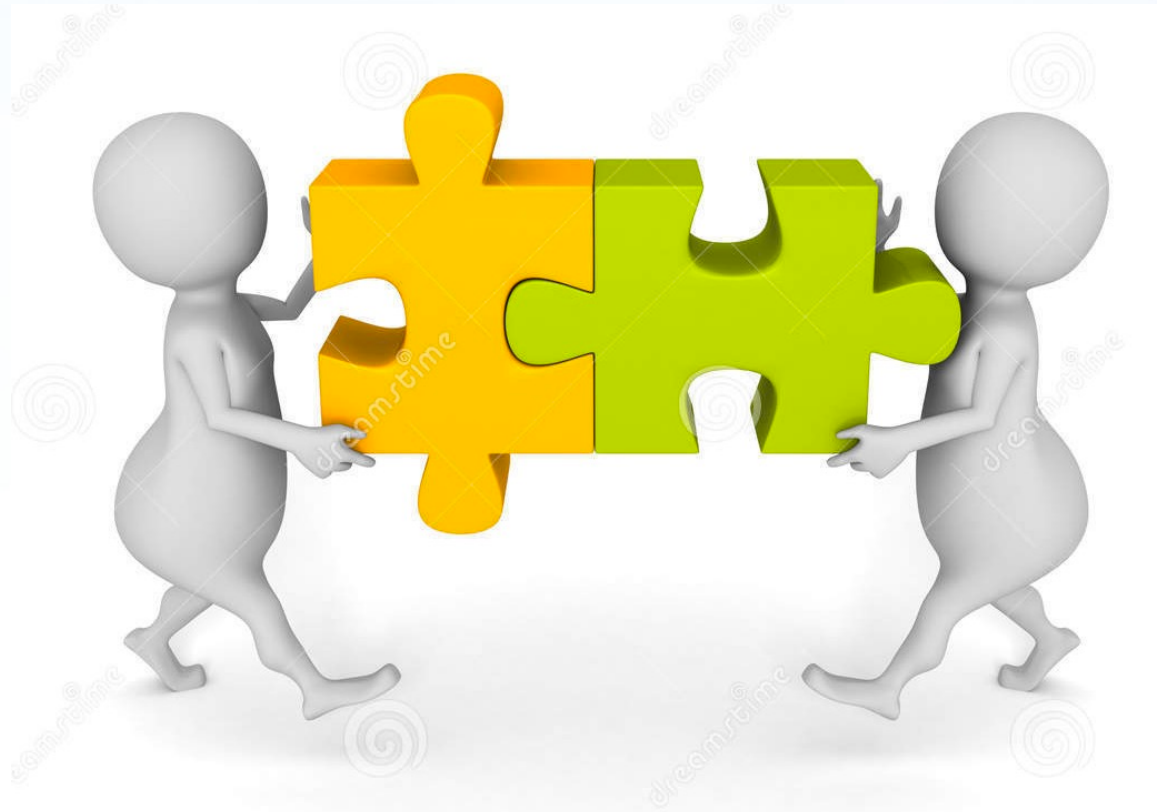
Publications

Download

News

Free Ideas

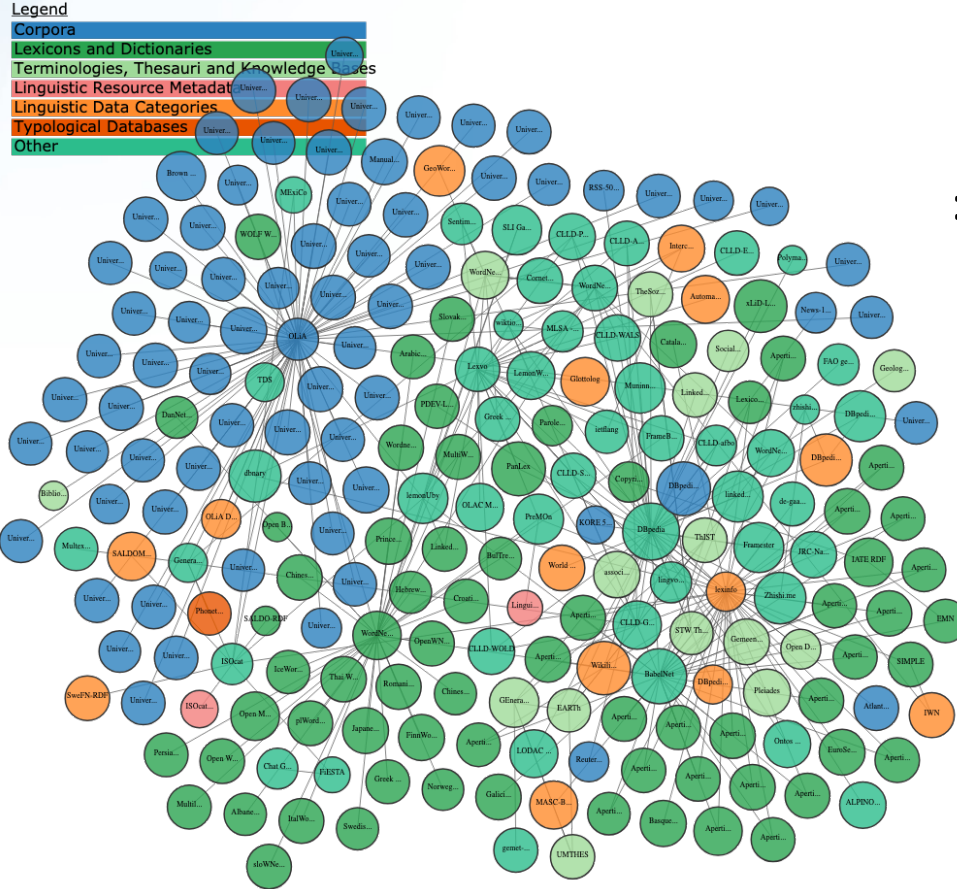
بقولكوا	قول	قال 1	قال 1	بقولكو	b/PROG_PART+A/IV 1S+qwi/IV+kw/IVSUF F_DO:2P	said;say;be_said; (for_example)_[CALIMA]	Verb	1	M	بننتا زي كعك العيد بتتفق بسرعة. العريس : والله انتو عيلة ما بشتريكووا بقمعة سيجارة , اخر اشي بقولكو اياه غروس مالبورو شو قلنوا ؟ الام : اسمع ولا , خاي ا
بقولك	قول	قال 1	قال 1	ببقولك	b/PROG_PART+y/IV 3MS+qwi/IV+k/IVSUF F_DO:2MS	said;say;be_said; (for_example)_[CALIMA]	Verb	3	M	زيون يا بقتل امي . مثال : عنا بتطلع بالتكسي . . . ايه ازمة . بتسال الشوفير شوف س ليش مازمة ؟ بقولك ما هو اليوم الخميس هيك . بتطلع السبت يا زلمة شو القص
بقولك	قول	قال 1	قال 1	ببقولك	b/PROG_PART+y/IV 3MS+qwi/IV+k/IVSUF F_DO:2MS	said;say;be_said; (for_example)_[CALIMA]	Verb	3	M	ير شوف س ليش مازمة ؟ بقولك ما هو اليوم الخميس هيك . بتطلع السبت يا زلمة شو القصه ليش مازمة ؟ بقولك يوم السبت عادة هيك , بتطلع الاثنين له له له شوه الأز
بقولك	قول	قال 1	قال 1	ببقولك	b/PROG_PART+y/IV 3MS+qwi/IV+k/IVSUF F_DO:2MS	said;say;be_said; (for_example)_[CALIMA]	Verb	3	M	مة شو القصه ليش مازمة ؟ بقولك يوم السبت عادة هيك , بتطلع الاثنين له له له شوه الأزمة شو في ؟ بقولك يوم الاثنين فش مجال دايم هيك . بتطلع الثلاثاء ول لي
بقولك	قول	قال 1	قال 1	ببقولك	b/PROG_PART+y/IV 3MS+qwi/IV+k/IVSUF F_DO:2MS	said;say;be_said; (for_example)_[CALIMA]	Verb	3	M	الثلاثاء ول ليش كل هالأزمة . يوم الثلاثاء بختي معروف فس نفس . بتطلع الخميس ليش هالقة الأزمة بقولك يوم الخميس دايم هيك , هذا يوم في الاسبوع طابير . بترو
بقولك	قول	قال 1	قال 1	ببقولك	b/PROG_PART+y/IV 3MS+qwi/IV+k/IVSUF F_DO:2MS	said;say;be_said; (for_example)_[CALIMA]	Verb	3	M	اط@@صنند@@ط@@ة@@منا. عنا بتطلع بالتكسي . . . ايه ازمة . بتسال الشوفير شوف س ليش مازمة ؟ بقولك ما هو اليوم الخميس هيك . بتطلع السبت يا زلمة شو القص



ربط بين المصادر اللغوية

شبكة عالمية من البيانات المعجمية المترابطة

The Linguistic Linked Open Data Cloud (LLODC)



التوجه العالمي الحديث للتعامل مع المصادر اللغوية:

- جهود بحثية مشتركة لإنشاء شبكة عالمية من البيانات المعجمية المترابطة.
- يتم تمثيل كل مدخلة ومفهوم وتعريف... في كل مصدر لغوي بطريقة (lemon) وربطها مع مقابلاتها في المصادر واللغات الأخرى.

قيد البحث والتطوير حالياً

بناء شبكة بيانات لغوية ضخمة للغة العربية تحتوي جميع المستويات التصريفية والاشتقاقية والدلالية والعامية.

Arabic Linked Data Cloud

ربط المعاجم من خلال ربط كل فرع (lemma) وتصريفاتها (inflections) لكل مدخلة في الـ 150 معجماً، ثم ربطها جميعاً بشبكات المصادر اللغوية الأجنبية.

Arabic Lemma Index

قائمة بجميع المدخلات العربية القديمة والحديثة والفصحى والعامية، وتحديد سماتها الصرفية ومعانيها وعلاقاتها الاشتقاقية.

فهم وربط لجميع المستويات اللغوية

خمس دئايء بوصل الكمر

CODA: القمر
POS_ar: اسم/قمر/اسم
POS_bw: Al/DET + qmr/NOUN
Gender: مذكر
Number: مفرد
Lemma: قَمَر1
....

CODA: دقائق
POS_ar: اسم/ادقايق
POS_bw: dqAyaq/NOUN
Gender: مؤنث
Number: جمع
Lemma: دَقِيقَة 1
....

جملة
عامية/فصحى

مستوى لغوي
تصريفي/اشتقاقي

مستوى
دلالة انطولوجية

المدخلة: قَمَر 1

(المصدر: قاعدة بيانات سما)
اللغة: فصحي حديثة
نوع المدخلة: اسم
تصريفات أخرى: قَمَر، قَمَر، قَمَر، قَمَر، قَمَر، قَمَر،
المعنى:
(AO 51587) moon ◦

المدخلة: دَقِيقَة 1

(المصدر: قاعدة بيانات سما)
اللغة: فصحي حديثة
صيغة المفرد المؤنث: دَقِيقَة
صيغة المثنى المؤنث: دَقِيقَتَيْنِ / دَقِيقَتَيْنِ /
تصريفات أخرى: دَقَائِقْ، دَقَائِقْ، دَقَائِقْ، دَقَائِقْ،
المعنى:
(AO_293582) minute ◦

جسم سماوي | جسم فلكي
Astronomical Body
A Physical object exists naturally in space

موجود مادي يتواجد بشكل طبيعي في الفضاء
example: النجم هو جسم سماوي متوهج وممتلئ ومضيء بذاته
293258 TypeOf: (physical object)

قمر | تابع satellite

جسم سماوي يدور حول كوكب آخر، لا يشمل ذلك الكواكب التي تدور حول شمس
example: نيل ارمنسترونغ هو اول رجل خطا على قمر الأرض
51587 TypeOf: (Astronomical Body)

دَقِيقَة

فترة زمنية تشكل جزءاً من ستين جزءاً متساوياً من الساعة
example: توجد 1436 دقيقة في اليوم الواحد
293582 TypeOf: (Interval)

(LLODC)

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 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. عُضُو الإِبْصَارِ فِي الْإِنْسَانِ وَالْحَيَوَانَ: لَهُ عَيْنَانِ كَعَيْنَيْ الصَّقْرِ - أَلَا إِنَّمَا الْعَيْنَانِ لِلْقَلْبِ رَائِدٌ ...
2. جَاسُوسٌ، "كَانَ عَيْنًا لِدَوْلَةٍ أجنبيَّةٍ . بَثَّ الْعَيُونُ : تَجَسَّسٌ، رَاقِبٌ - فَلَانٌ عَيْنٌ عَلَى فَلَانٍ : نَاطِرٌ عَلَيْهِ
3. أَجُودُ كُلِّ شَيْءٍ وَأَحْسَنُهُ وَنَفِيسُهُ: عَيُونُ الْفَنِّ.
4. حَارِسٌ: فَلَانٌ عَيْنٌ عَلَى الْمَكَانِ.
5. الْحَاضِرُ مِنْ كُلِّ شَيْءٍ أَصْبَحَ أَثَرًا بَعْدَ عَيْنٍ ...
6. عَيْنُ الْمَاءِ:- يَنْبُوعُهُ، تُحَلِّقُ الطَّيُورُ فَوْقَ عَيُونِ الْمَاءِ
7. عَيْنُ الشَّيْءِ:- نَفْسُهُ، ذَاتُهُ (تَسْتَعْمَلُ لِلتَّوَكِيدِ): جَاءَ الْقَوْمُ أَعْيُنَهُمْ - كُنَّا فِي الْمَكَانِ عَيْنَهُ.
8. عَيْنُ الْعَقْلِ:- قُدْرَةُ ذَهْنِيَّةٍ مَوْرُوثَةٍ عَلَى التَّخْيِيلِ وَتَذَكُّرِ الْأَحْدَاثِ.
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.

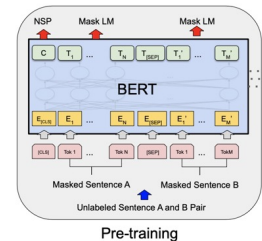
❖ Arabic context-gloss pairs Dataset (167k)

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

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



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

شكرا لاهتمامكم

مصطفى جرار - جامعة بيرزيت
mjarrar@birzeit.edu