استخراج المترادفات آلياً (بيانات وخوارزميات)

https://sina.birzeit.edu/synonyms

لقد طُوِّرَت عدة خوارزميات وأدوات برمجية وبيانات من أجل استخراج المترادفات العربية آلياً. إلا أن الفكرة الجديدة في هذا العمل لا تقتصر على الجانب الحاسوبي فقط، بل وعلى فكرة التعامل مع المترادفات كعلاقة نسبية (fuzzy relationship). قمنا أولاً بإجراء تجربة علمية عمل فيها أربعة لغويين على توسيم ثلاث آلاف مترادفة، ووُسِمَت كل مترادفة بعلامة تدل على مدى قوتها الترادفية (fuzzy value). واستُخدِمَت هذه المدونة لفحص دقة الخوارزميات التي أُجرِيَ تطويرها وتبين أن هذه الخوارزميات قد قاربت دقة وسلوك اللغويين، إذ تقوم الخوارزمية بتحليل القواميس اللغوية المتعددة اللغات (translation pairs) واستنتاج علاقات ترادف جديدة.

يمكن للمستخدم، لغرض التجريب، إدخال كلمة أو أكثر (بالعربية أو الإنجليزية) وستقوم الخوارزمية باقتراح مجموعة مترادفات، وإعطاء قيمة (fuzzy value) لمدى ملائمة وقوة هذا الترادف.

- تُستخدم الأدوات أيضاً لتقييم دقة المترادفات وليس فقط لاستخراجها. فإذا أُدْخِلَت مجموعة مترادفات فإنه يتم إعطاء علامة تعبّر عن مدى دقة الترادف لكل كلمة.
- يمكن استخدام الأدوات لاستخراج المترادفات لأية لغة (في حال أُجْرِيَ تحميل معاجم لتلك اللغة). وتجدر الإشارة إلى أن الخوارزمية استُخْدِمَت لاستخراج مكنز مترادفات خاص باللغة الويلزية (Welsh) بنجاح، ونُشِرَت ورقة حول الموضوع (انظر الأوراق أدناه).
- جميع الأدوات والخوارزميات والبيانات مفتوحة المصدر ومتاحة للتجريب (الرابط أعلاه).
- نُشِرَت عدة أوراق علمية وصفت المشروع من الناحية البحثية (حوسبة اللغة).

أوراق علمية ذات علاقة بالمترادفات

للمزيد يمكن الضغط على الرابط بلون (تركواز)

Sana Ghanem, Mustafa Jarrar, Radi Jarrar, Ibrahim Bounhas: **A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms.** The 12th International Global Wordnet Conference (GWC2023), Global Wordnet Association. (pp.). San Sebastian, Spain, 2023 * الورقة تصف البيانات والخوارزميات لاستخراج المترادفات العربية
paper: http://www.jarrar.info/publications/GJJB23.pdf
slides: http://www.jarrar.info/Talks/GWN2023_synonyms.pdf

Nouran Khallaf, Elin Arfon, Mo El-Haj, Jonathan Morris, Dawn Knight, Paul Rayson, Tymaa Hammouda, Mustafa Jarrar: Open-Source Thesaurus Development for Under-Resourced Languages: a Welsh Case Study. The Language, Data and Knowledge (LDK 2023) Conference Vienna, Austri, 2023. ورقة تم فيها تجريب الخوارزمية لاستخراج مترادفات بلغة الـ Welsh

Eman Naser-Karajah, Nabil Arman, Mustafa Jarrar: Current Trends and Approaches in Synonyms Extraction: Potential Adaptation to Arabic. In Proceedings of the 2021 International Conference on Information Technology (ICIT). PP 748--755, Association for Computational Linguistics. pp. 428-434, IEEE. 2021 ورقة مسحية لأحدث تقنيات استخراج المترادفات

Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, Khaled Shaalan: Extracting Synonyms from Bilingual Dictionaries. The 11th International Global Wordnet Conference (GWC2021), Global Wordnet Association. (pp. 215-222). Pretoria, South Africa, 2021 ورقة تقترح خوارزمية لاستخراج المترادفات من المعاجم ثنائية اللغة

مثال (screenshot) لاستخراج المترادفات آلياً

https://sina.birzeit.edu/synonyms

أعطيت الخوارزمية كلمتين (شارع و طريق)، وتم استرجاع عشرات المترادفات بالعربية والانجليزية لهما، وكل مترادفة تم إعطائها قيمة تدل على قوة الترادف

SinaLab

News Team Resources

Synonyms Generator

A dataset and source code for Synonyms Generator

Version: 1.0 (updated on 15/12/2022)

An algorithm to extend a set of synonyms with more synonyms. Given a set of synonyms, the algorithm builds a graph (using many dictionaries) and returns a set of candidate synonyms, each with a fuzzy value to indicate how much it is likely to be a synonym. The more synonyms in the input, the more accurate the candidate synonyms. We trained the fuzzy model using a dataset we built manually (500 synsets, with 3K candidate synonyms by four linguists). Please read this article for the details. Try the service (type synonym seperated by | or , or ،):

طريق | شارع

Extend Evaluate

street 61% , road 61% , way 30% , via 30% , track 30% , thoroughfare 30% , route 30% , roadway 30% , road way 30% , ride 30% , path 30% , highway 30% , boulevard 30% , avenue 30% , Roadway 30% , Lane 30% , Boulevard 30% , Avenue 30%

مَمَرّ 61% , مَسْلَكٌ 61% , مور 61% , صِرَاطٌ 61% , سَبِيلٌ 61% , سَنَن 61% , زُقَاقٌ 61% , دَرْب 61% , وَجْه 30% , نَهْج 30% , نَمَطٌ 30% , نَهْج 30% , مِنْهَج 30% , مِنهاج 30% , مَمَر 30% , مَسَار 30% , مَرصَد 30% , مَذْهَبٌ 30% , مَدْرَج 30% , مَحَجَّة 30% , مصطبة السور 30% , مرقب 30% , مذهب 30% , مجال 30% , قسم 30% , طَرِيقَةٌ 30% , طَرِيقٌ 30% , طرز 30% , صُورَة 30% , شَارِعٌ 30% , سِيرَة 30% , سِكَّةٌ 30% , سِرَاط 30% , سُنَن 30% , سُلوك 30% , سَبِيْل 30% , سيرة 30% , سلوك 30% , زَنقة 30% , رُواق 30% , ربع 30% , جَادَّة 30% , أُسْلُوبٌ 30%

- Dataset And Downloads

The dataset is a set of 500 synsets (extracted from the Arabic Wordnet). Each synset is enriched with a list of *candidate* synonyms. The total number is 3K candidates. Each candidate synonym is then annotated with a fuzzy value by four linguists (in parallel). The dataset is important for understanding how much linguists (dis/)agree on synonymy (which we found RMSE: 32% and MAE: 27%). In addition, we used the dataset as a baseline to evaluate our algorithm. See the scoring guidelines, figures, and details in section 3.

License: MIT

Download: Github.Synonyms

Please email Prof. Mustafa Jarrar (mjarrar AT birzeit.edu) if you have any question.