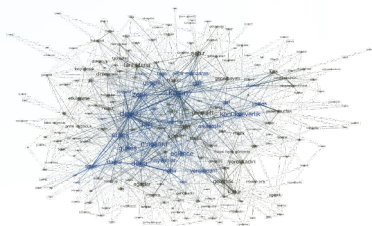# ArabGlossBERT

Fine-Tuning BERT on Context-Gloss Pairs for Word Sense Disambiguation

Moustafa Al-Hajj
Lebanese University
Lebanon

Mustafa Jarrar
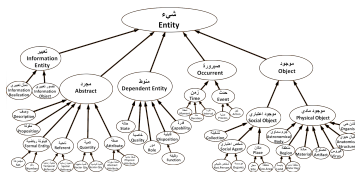Birzeit University
Palestine

Lexical Resources at Birzeit University

Lexicographic Database

150 lexicons
Largest 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

ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD

Moustafa Al-Hajj
Lebanese University
Beirut, Lebanon
moustafa.alhajj@ul.edu.lb

Mustafa Jarrar*
Birzeit University
Birzeit, Palestine
mjarrar@birzeit.edu

Abstract

Using pre-trained transformer models such as BERT has proven to be effective in many NLP tasks. This paper presents our work to fine-tune BERT models for Arabic Word Sense Disambiguation (WSD). We treated the WSD task as a sentence-pair binary classification task. First, we constructed a dataset of labeled Arabic context-gloss pairs (~167k pairs) we extracted from the Arabic Ontology and the large lexicographic database available at Birzeit University. Each pair was labeled as *True* or *False* and target words in each context were identified and annotated. Second, we used this dataset for fine-tuning three pre-trained Arabic BERT models. Third, we experimented the use of different supervised signals used to emphasize target words in context. Our experiments achieved promising results (accuracy of 84%) although we used a large set of senses in the experiment.

1 Introduction

Word Sense Disambiguation (WSD) aims to determine which sense (*i.e.* meaning) a word may denote in a given context. This is a challenging social media mining, Named-Entity Recognition, word sense disambiguation, topic classification and summarization, among others.

A gloss is a short dictionary definition describing one sense of a lemma or lexical entry (Jarrar, 2006, 2005). A context is an example sentence in which the lemma or one of its inflections (*i.e.* the target word) appears. In this paper, we aim to fine-tune Arabic models for Arabic WSD. Given a target word in a context and a set of glosses, we will fine-tune BERT models to decide which gloss is the correct sense of the target word. To do that, we converted the WSD task into a BERT sentence-pair binary classification task similar to (Huang et al., 2019; Yap et al., 2020; Blevins and Zettlemoyer, 2020). Thus, BERT is fine-tuned on a set of context-gloss pairs, where each pair is labeled as *True* or *False* to specify whether or not the gloss is the sense of the target word. In this way, the WSD task is converted into a sentence-pair classification task.

One of the main challenges for fine-tuning BERT for Arabic WSD is that Arabic is a low-resourced language and that there are no proper labeled context-gloss datasets available.

To overcome this challenge, we collected a rela-

The Word Sense Disambiguation (WSD) Task

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

<div dir="rtl">

قصيدة من <u>عيون</u> الشعر

</div>

The problem

Set of senses

<div dir="rtl">

1. **عُضو الإبصار** في الإنسان والحيوان: له عينان كعَيْنَيْ الصقر ـ ألا إنّما العينان للقلب رائدٌ ...
2. **جاسوس،** "كان عيئًا لدولةٍ أجنبيَّة . بثَّ العيون : تجسَّس، راقب ـ فلانٌ عَيْن على فلان : ناظر عليه
3. **أجود كلّ شيء** وأحسنه ونفيسُه: عيونُ الفنِّ.
4. **حارس:** فلان عين على المكان.
5. **الحاضر من كل شيء** أصبح أثرًا بعد عين ...
6. **عَيْنُ الماء:**ـ ينبوعه، تُحلِّق الطَّيورُ فوق عيون الماء
7. **عَيْن الشَّيء:**ـ نفسه، ذاته (تستعمل للتوكيد): جاء القوم أعينهم ـ كنّا في المكان عينه.
8. **عَيْن العقل:**ـ قدرة ذهنيَّة مَوْروثة على التخيُّل وتذكَّر الأحداث.
9. .....

</div>

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

Pre-training
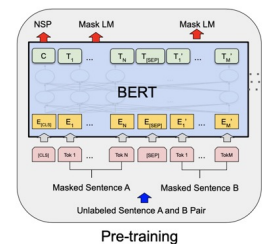
Related Work

BERT-based:

Used context-gloss pairs in fine-tuning (Huang et al., 2019; Yap et al., 2020; Blevins and Zettlemoyer, 2020). El-Razzaz et al. (2021) used a small dataset, but the experiment is not reliable.

Used markers to emphasize target words in context-gloss Huang et al. (2019); Botha et al. (2020); Lei et al. (2017);

Static Embeddings (related to Arabic WSD)

Sense vectors (Laatar et al., 2017), Stem2Vec and Sense2Vec(Alkhatlan et al., 2018), Lemma2Vec (Al-Hajj and Jarrar, 2021), Word Sense Induction (Alian and Awajan, 2020), or using fastText (Logacheva et al., 2020). Elayeb (2019) reviewed Arabic WSD approaches until 2018.

Constructing a dataset of context-gloss pairs

- Extracted (60K pairs) from the Arabic Ontology and the ~400K dictionary definitions available at Birzeit University. All are labeled as *True*.

- Generated 107k *False* pairs from the *True* pairs.

➢ Annotated target words in context

https://ontology.birzeit.edu
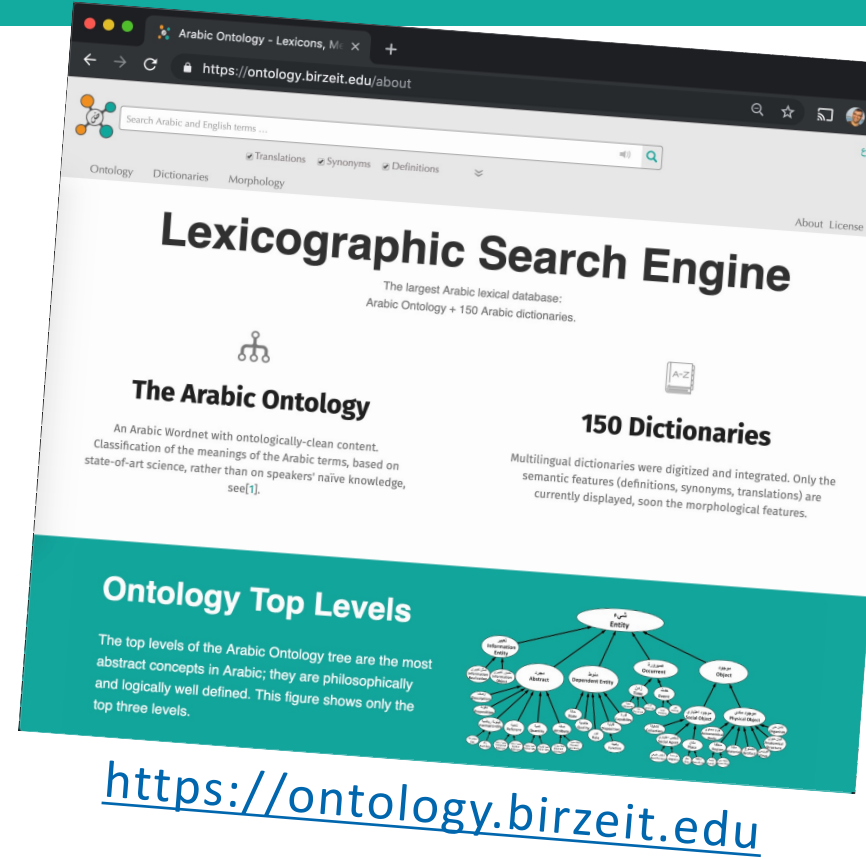
Constructing a dataset of context-gloss pairs

Statistics:

	count
Unique Lemmas (undiacritized)	26169
Avg glosses per Lemmas	1.25
Unique Glosses	32839
Unique Contexts	60272
Avg context per gloss	1.83
True context-gloss pairs	60323
False context-gloss pairs	106884
Total True and False pairs	167207

Training and Test Datasets
- every context selected in the test set should not be selected in the training set;
- every gloss should be selected in both the training and the test sets.

Datasets	Pairs	Count	Total
Training	True pairs	55,585	
	False pairs	96,450	152,035
Test	True pairs	4,738	
	False pairs	10,434	15,172
		Total	167,207

Download: https://ontology.birzeit.edu/downloads

Constructing a dataset of context-gloss pairs

- Annotating Target Words in Context

  - Used four methods in parallel
    - Substring
    - Character-level cosine similarity
    - Levenshtein distance
    - Lemmatization

  - All results were **validated manually**

Lemma	**عين**
Gloss	أجود كل شيء وأحسنه ونفيسه
Context	قصيدة من **عيون** الشعر

- Used four variations of tagging target words

  supervised signals used to emphasize target words in context during the BERT fine-tuning

Variation 1	[CLS] قصيدة من عيون الشعر [SEP] أجود كل شيء وأحسنه ونفيسه [SEP]
Variation 2	[CLS] قصيدة من 'عيون' الشعر [SEP] عيون: أجود كل شيء وأحسنه ونفيسه [SEP]
Variation 3	[CLS] قصيدة من [UNUSED0]عيون[UNUSED0] الشعر [SEP] عيون: أجود كل شيء وأحسنه ونفيسه [SEP]
Variation 4	[CLS] قصيدة من [UNUSED0]عيون[UNUSED1] الشعر [SEP] عيون: أجود كل شيء وأحسنه ونفيسه [SEP]

❖ Fine-tuned three Arabic pre-trained BERT models

The WSD task is converted into **binary sequence-pair classification task**

[CLS] قصيدة من عيون الشعر [SEP] أجود كل شيء وأحسنه ونفيسه [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 وأحسنه ونفيسه [SEP]	False

❖ Results

Model		True	False	Accuracy
AraBERTv02	Precision	81	85	84
	Recall	66	93	
	F1-score	72	89	
CAMeLBERT	Precision	77	83	82
	Recall	60	92	
	F1-score	67	87	
QARiB	Precision	73	82	80
	Recall	58	90	
	F1-score	65	86	

Fine-Tuning

Tagging target words while fine-tuning

To emphasize target words during the BERT fine-tuning

Tested 4 variations of tags

❖ **Results** (with AraBERTv02)

Variation		True	False	Accuracy
Variation 1 No signal	Precision	80	85	
	Recall	64	92	83
	F1-score	71	88	
Variation 3 UNUSED0	Precision	81	85	
	Recall	64	93	84
	F1-score	71	89	
Variation 4 UNUSED0,1	Precision	81	85	
	Recall	64	93	84
	F1-score	71	89	

Summary

Converted: **WSD** task into ***binary sequence-pair classification***

Constructed Dataset: labeled Arabic context-gloss pairs (167K)

Annotated: Target words in context, to emphasize them

Fine-tuned: BERT: three models (84% accuracy)

Thank You

Mustafa Jarrar
mjarrar@birzeit.edu

References

Moustafa Al-Hajj and Mustafa Jarrar. 2021. LUBZU at SemEval-2021 task 2: Word2Vec and Lemma2Vec performance in Arabic word-incontext disambiguation. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 748–755, ACL.

Diana Alhafi, Anton Deik, and Mustafa Jarrar. 2019. Usability evaluation of lexicographic e-services. In The 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), pages 1–7. IEEE.

Hamzeh Amayreh, Mohammad Dwaikat, and Mustafa Jarrar. 2019. Lexicon digitization-a framework for structuring, normalizing and cleaning lexical entries. Technical Report, Birzeit University.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta CavalliSforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab worlds. Commun. ACM, 64(4):72–81.

Mustafa Jarrar. 2005. Towards Methodological Principles for Ontology Engineering. Ph.D. thesis, Vrije Universiteit Brussel.

Mustafa Jarrar. 2006. Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In Proceedings of the 15th international conference on World Wide Web (WWW2006), pages 497–503. ACM Press, New York, NY.

Mustafa Jarrar. 2011. Building a formal arabic ontology (invited paper). In Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Alecso,Arab League.

Mustafa Jarrar. 2021. The arabic ontology - an arabic wordnet with ontologically clean content. Applied Ontology Journal, 16(1):1–26.

Mustafa Jarrar and Hamzeh Amayreh. 2019. An arabicmultilingual database with a lexicographic search engine. In The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019), volume 11608 of LNCS, pages 234– 246. Springer.

Mustafa Jarrar, Hamzeh Amayreh, and John P. McCrae. 2019. Representing arabic lexicons in lemon - a preliminary study. In The 2nd Conference on Language, Data and Knowledge (LDK 2019), volume 2402, pages 29–33. CEUR.

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, (3).

Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, and Khaled Shaalan. 2021. Extracting synonyms from bilingual dictionaries. In The 11th International Global Wordnet Conference (GWC2021), pages 215–222. Global Wordnet Association.

Mustafa Jarrar, Fadi Zaraket, Rami Asia, and Hamzeh Amayreh. 2018. Diacritic-based matching of arabic words. ACM Asian and Low-Resource Language Information Processing, 18(2):10:1–10:21.

Eman Naser-Karajah, Nabil Arman, and Mustafa Jarrar. 2021. Current trends and approaches in synonyms extraction: Potential adaptation to arabic. In 2021 International Conference on Information Technology (ICIT), pages 428–434, Amman, Jordan. Association for Computational Linguistics.