Enrichment of Arabic WordNet Using Machine Translation and Transformers

Mohamed Dia Eddine Souci [1], Younes Cherifi [1], Lamia Berkani [1], Mohamed Seghir Hadj Ameur [2], Ahmed Guessoum [2]

[1] Faculty of Informatics, USTHB University, Bab Ezzouar, Algiers , Algeria
[2] Higher National School of Artificial Intelligence (ENSIA), Sidi Abdellah, Algiers, Algeria

lberkani@usthb.dz; {mohamed.hadj.ameur; ahmed.guessoum}@ensia.edu.dz

Abstract

This article aims to enhance Arabic WordNet (AWN) by exploiting the current version of the Princeton WordNet (PWN) and Deep Learning (DL) techniques, known for their effectiveness in machine translation. We aim to improve the coverage and quality of AWN by adding new Synsets, integrating definitions and examples, and validating semantic relationships. The contribution can be summarized in three aspects: (1) utilizing multiple translation systems to translate PWN resources and web-extracted data into Arabic; (2) employing Transformers, a highly effective deep learning technique, to refine the outcomes from the first step; and (3) developing a web portal that enables users to visualize proposed updates and facilitates validation by human experts. The results from the evaluation of a random sample of 1,000 Synsets taken from the candidate pool for enrichment are highly promising, with a manual validation accuracy of 75.4%. With such a validation accuracy, the potential would be to get almost 68,000 Synsets correct out of the 90,127 candidates produced by our approach.

1 Introduction

In most of Natural Language Processing applications, the effective handling of semantics relies heavily on the presence of lexical-semantic resources. Among these, WordNets stand out as crucial ones. They can be used in tasks like text classification, information retrieval, and semantic analysis (Morato et al., 2004). WordNets are lexical databases that organize language words based on their meanings. These databases are built on a foundational structure known as the WordNet backbone, which encompasses a hierarchical arrangement of conceptual categories referred to as a taxonomy. Each of these concepts is represented by a set of lemmas (words) that share identical meanings, forming what is known as a Synset. This framework essentially forms a semantic network where words are connected to their corresponding concepts.

The challenge lies in constructing these valuable databases. Although the initial WordNet, known as the Princeton WordNet (Fellbaum, 1998), was meticulously constructed for the English language by linguistics experts and established itself as a standard reference, many other languages lack the resources required to create comparably comprehensive WordNets for various applications. The manual construction and expansion of these WordNets are laborious and resource-intensive tasks. Thus, researchers are striving to devise methods to automate these processes and minimize human involvement.

Our work is primarily based on leveraging PWN, because of its substantial size and extensive coverage, in the enrichment of the AWN. Initially, our approach involves gathering all the grouping lemmas (names) of Synsets along with their existing definitions from PWN. Subsequently, we translated them from English to Arabic using multiple translation engines. To further enrich AWN, recognizing that a majority of PWN Synsets lack examples, we have devised a method to extract a set of relevant examples for each Synset from the web, based on its context. By so doing, we have constructed potential Arabic Synsets, maintaining the same relationships present in PWN. We have obviously taken care of preserving the same hierarchical structure as well as the mapping from PWN. Once these candidate Synsets are generated, a validation process becomes crucial. This step involves manual validation by human experts through a web-based validation portal. Finally, we have extracted the outcomes of the validation step and compiled the newly enriched AWN by incorporating the validated candidate Synsets.

2 Related work

AWN is in a constant state of evolution and expansion, with numerous research efforts aimed at enriching and improving it since its inception in 2006 (Black, et al., 2006).

The initial attempt to enhance AWN was conducted by Alkhalifa and Rodríguez (2018). This work involved utilizing the Wikipedia encyclopedia for the automatic extraction of Arabic named entities that had English equivalents in PWN. Boudabous et al (2013) proposed a linguistic method based on two phases. The first one defines morpho-lexical patterns using a corpus developed from Arabic Wikipedia. The second phase uses patterns to extract new semantic relations from AWN entities. Abouenour et al. (2016) presented an enrichment of AWN targeting three types of content required by Arabic Q-A systems: (1) enrichment of instances or named entities; (2) enrichment of verbs and nouns by extending the list of verb senses and refining the hyponymy relationship between AWN noun Synsets; and (3) enrichment of broken (i.e. irregular) plurals, a class of plural forms that is widely used and precisely defined in Arabic. Hadj Ameur et al. (2017) proposed an automated approach to enrich WordNets sharing the same structural framework as PWN and whose existing Synsets (concepts) are mapped onto the PWN concepts. The authors employed resource-based methods, including dictionaries and ontologies, along with corpus-based methods to extract unambiguous Arabic lemmas and lexical relations. Subsequently, they translated the lemmas into English and paired them with their corresponding PWN lemmas, generating a set of candidate Synsets. Each candidate was assigned a score using a set of features, and these feature parameters were optimized using a suitable metaheuristic. Finally, vocalization and usage examples were provided for each candidate and added to AWN. Batita and Zrigui (2018) focused on enriching antonymy relationships in AWN. Lam et al. (2014) proposed approaches for generating WordNet Synsets for both resource-rich and resource-poor languages, using publicly available WordNets, an automatic translator and/or a single bilingual dictionary. Al Tarouti and Kalita (2016) introduced a novel enrichment approach to enhance the conventional translation approach by utilizing word vectorization (Word Embeddings). This approach involves constructing an initial WordNet in the target language T and enriching it using the approach presented in (Lam et al. 2014). Subsequently, using the Word2vec algorithm, the authors generated word vectors from an existing corpus. These vectors were then used to filter words that belong to each generated Synset, retaining only word pairs with the highest cosine similarity. Utilizing the same similarity measure and existing WordNets such as PWN, a similarity threshold was computed between pairs of synonymous words and semantically related words. This threshold was then used to validate candidate Synsets. Batita and Zrigui (2019) provide a case study on the updates of AWN and the development of its contents, focusing on the relations that have been added to the extended version.

3 Design

In this section, we will explore the intricacies of the approach we have designed to enhance AWN. The process involves several key stages, beginning with the alignment of English and Arabic WordNets and progressing through the extraction of essential elements, including definitions, examples and the translation of words. Ultimately, we will elucidate the final step in this process: the integration of these enriched elements into AWN.

3.1 Alignment with PWN

In order to enrich AWN, we will adopt an alignment-based approach with the Princeton WordNet, which contains the most extensive set of Synsets among all WordNets, with a total of 117,659 Synsets, compared to the 11,269 Synsets of AWN. We establish selection criteria for words extracted from PWN. An English word will be considered a candidate for enrichment if it lacks a corresponding translation in Arabic within AWN, as illustrated in Figure 1. In this way we will have obtained a new, reduced subset of PWN which contains only the Synsets that will be considered in the AWN enrichment operation.

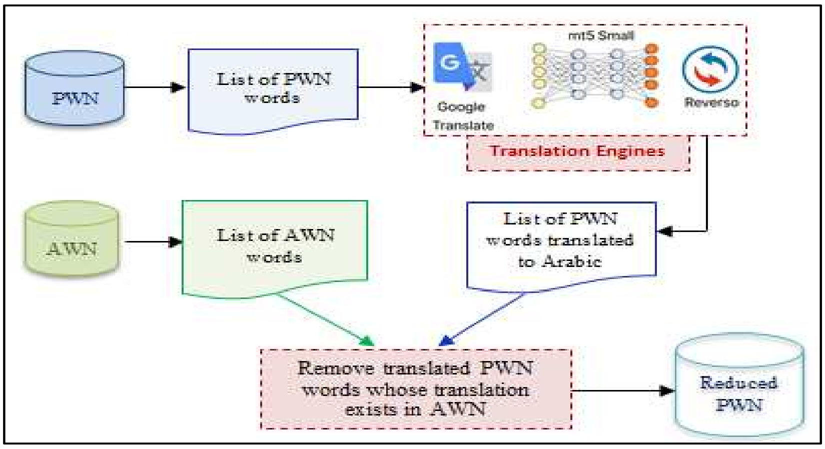


Figure 1: PWN-AWN alignment schema

3.2 Extraction of definitions

Definitions contextualize the word by placing it within its semantic context. We leveraged definitions in our enrichment approach to more accurately determine the meaning of translated words and eliminate any possible ambiguity. This step in our approach involves gathering definitions of words from each Synset that will be integrated into AWN. To accomplish this, we needed a lexical resource that would allow us to collect data automatically. However, we did not find such a resource for the Arabic language. To address this challenge, we decided to use the definitions already present in PWN and directly translate them into Arabic. We followed the steps illustrated in Figure 2 and described below it.

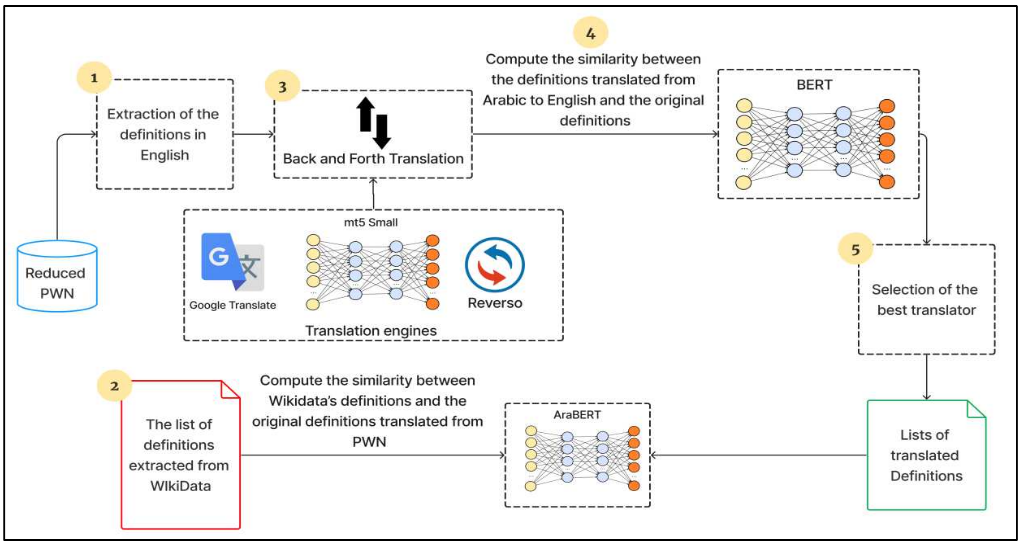

1
Extraction of the definitions in English
3
Back and Forth Translation
4
Compute the similarity between the definitions translated from Arabic to English and the original definitions
BERT
Reduced PWN
mt5 Small
Google Translate
Reverso
Translation engines
5
Selection of the best translator
2
The list of definitions extracted from WIkiData
Compute the similarity between Wikidata's definitions and the original definitions translated from PWN
AraBERT
Lists of translated Definitions

Figure 2: Extraction of definitions

1) We proceeded by retrieving, for each Synset in PWN, its English definition. Then we translated this into Arabic using various translation systems, including Google Translator, Reverso, and our custom translation model developed using the Transformer MT5.

2) Additionally, we extracted other definitions from Wikidata. However, after manually analyzing some of these definitions, we observed that the majority of them lacked precision in terms of the original context of the Synset. Nevertheless, we opted to retain them as candidate definitions, with the intention of getting them through automated evaluation at a later stage.

3) Following the compilation of lists of candidate definitions, we employed the Back-And-Forth Translation logic to compare the translated definitions with the original definitions and evaluate their similarities to determine the

most effective translation system among the aforementioned three.

4) To assess the similarity while considering the context of the definitions, we utilized the BERT language model to extract Word Embeddings and calculate cosine similarity between the translations of the definitions and their originals.

5) By comparing the original definitions with the translated candidate definitions, we selected the translation that maximized the similarity. This way we made sure that the translated definitions are as close as possible to the original definitions, while preserving context and meaning.

6) Finally, we generated a list of candidate definitions that provides a range of translated definitions tailored to context, accurately representing the meanings of the Synsets.

3.3 Word Translation

To generate the list of Arabic words translated from the names of English Synsets, we followed the same steps as those applied for the extraction of definitions (see Figure 3):

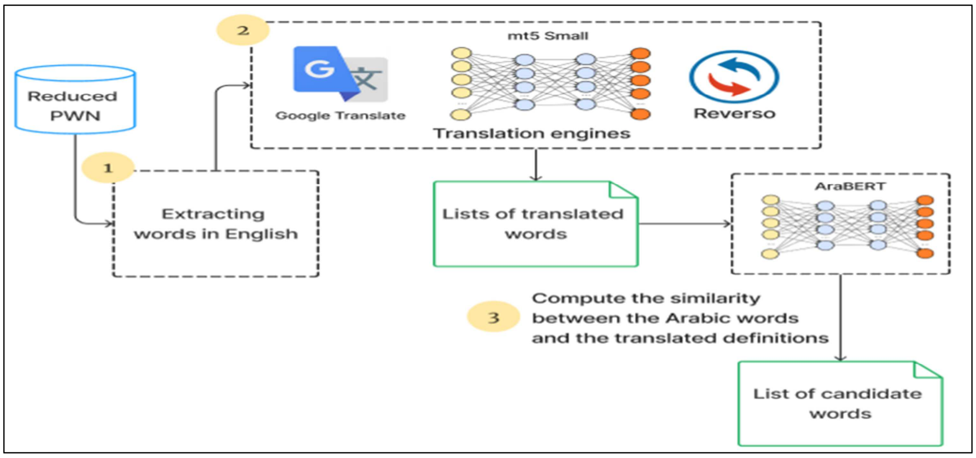

Reduced PWN
1
Extracting words in English
2
mt5 Small
Google Translate
Reverso
Translation engines
Lists of translated words
AraBERT
3
Compute the similarity between the Arabic words and the translated definitions
List of candidate words

Figure 3: Schema representing the process followed to translate and select words

The steps followed for the translation of words are described as follows:

1) First, we extracted words from the names of the Synsets available in PWN.

2) Similar to the approach used for definitions, we found it preferable to utilize a variety of translation tools such as Google Translate and Reverso Translate to obtain multiple candidate translations for each Synset. This was necessary because these translators do not consider the context of the word during translation.

3) After obtaining the translations of each word from the PWN Synsets, we evaluated their similarities with the translated Arabic definitions using the AraBERT model and the cosine similarity measure. This step enabled us to retain only the candidate word translations that exhibited maximum contextual similarity.

4) Finally, we generated a list of translated words by selecting the best candidates that correspond to the different meanings and contexts of the Synsets.

3.4 Extraction of Examples

Examples are highly valuable in WordNets as they provide concrete illustrations of word usages, associated senses, and the context in which a word would appear. To maximize the number of examples for each Synset, we followed the steps outlined below, as illustrated in Figure 5.

We utilized Web Scraping techniques to extract relevant data from the Reverso website, which provides examples covering various contexts and uses for most words in the Arabic language. The extracted data were saved in a text file to be used in the following steps. However, we found that not all the extracted examples were correct, i.e. some of them contained Latin characters or were simply in English. Therefore, a preprocessing was performed to eliminate these incorrect examples using a model we trained by fine-tuning the AraBERT model.

After generating the final list of examples for each Synset, we evaluated the similarity of each example with the previously chosen Arabic definition in the first step using the AraBERT model. Hence, by selecting examples that exhibited maximum similarity with the Arabic definition, we ensured that we chose an example closely aligned with the specific meaning and context of the Synset.

3.5 Validation Module

After generating a final list of candidates for each Synset (its translation, definition, and a list of contextual examples), this list was submitted to a web validation platform. This platform enables human experts to manually review each Synset. Experts could confirm the accuracy of the models used, identify potential errors or inconsistencies in the proposed Synsets, and make corrections.

3.6 Insertion into AWN

The final phase of our approach will focus in the future on generating the XML file containing the newly validated Synsets for integration into the current version of AWN, which is available on the GlobalWordNet portal.

We are optimistic about the acceptance of the vast majority of the generated Synsets since the validation platform will allow human experts to correct any potential errors that may be detected in terms of translation or inappropriate examples or definitions for a given Synset word. We expect that the number of retained words after filtering may be reduced in cases where the translation of the word and/or definition is rejected. If the semantic relations of a translated Synset are validated, we will include them in the XML file while maintaining the same hierarchical structure as that of PWN. However, if any semantic relation is not valid, we will add it as an independent Synset, meaning that it will not be linked to other Synsets in the hierarchy of the newly enriched AWN.

4 Experiment

This section covers the technical aspects related to the development of our approach, including the tools used. It also provides illustrations of each step of our approach. Finally, it will present a summary of the results obtained, along with an analysis and discussion.

4.1 Results of word extraction

To accomplish this task, we have made use of the WordNet library from the nltk.corpus package. We employed the wordnet.all_synsets () function, which returns a list containing all 117,659 Synsets from PWN 3.0. Subsequently, for each Synset, we retrieved its identifier using the synset.name () function. Each identifier has the format illustrated in Figure 4:

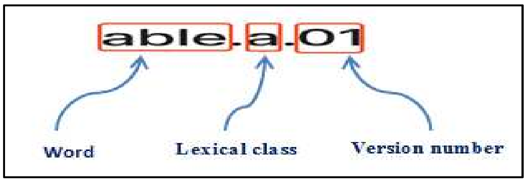
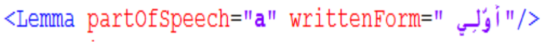


Figure 4: Synset ID Format

Finally, using Python programming we retrieve the word as the first part of the Synset ID.

We followed the same procedure for AWN, which contains 60,157 Synsets in its second version, "arb2-lmf," available for download from GlobalWordNet [1] website. This version of AWN is provided in XML format. We extracted words from the "writtenForm" attribute of the <Lemma> tag, as follows:



```
<Lemma partOfSpeech="a" writtenForm=" أوّلي "/>
```

Therefore, after retrieving the two lists of words, we retained only the words from PWN that do not have translations in AWN. Table 1 represents the number of extracted words:

WordNet	Number of extracted words
PWN	117,659
AWN	19,978
Reduced PWN	90,127

Table 1: Number of extracted words.

4.2 Choosing the best translation system

As previously mentioned, we used the Back-And-Forth Translation logic to select the translation system that provides the best Arabic translations. To achieve this, we opted for employing the original definitions from PWN as a reference for this task. Initially, we translated each definition from English to Arabic using three translation systems: Google Translate, Reverso, and our MT5 model.

Next, we back-translated the produced (Arabic) results into the English language. Subsequently, we employed cosine similarity to compare these back-translated definitions with the original English definitions. The translation

[1] http://globalwordnet.org/

system that yielded translations most closely aligned with the meaning and context of the original English definitions was chosen as the optimal system for translation. This approach ensures that the selected translation system produces as accurate and contextually relevant translations as possible for the enrichment of AWN.

In the following, we present two graphs that highlight the similarities between different translated definitions and the original definitions. The first graph in Figure 5 illustrates the similarities of a sample of 100 definitions translated with Google Translate and the MT5 model. The second graph in Figure 6 shows the similarities of the same sample of definitions translated with Google Translate and Reverso.

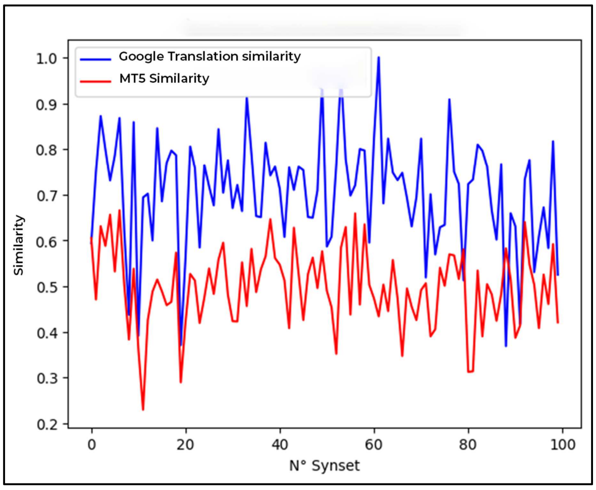


Figure 5: Comparison of translation similarities between Google Translate and MT5

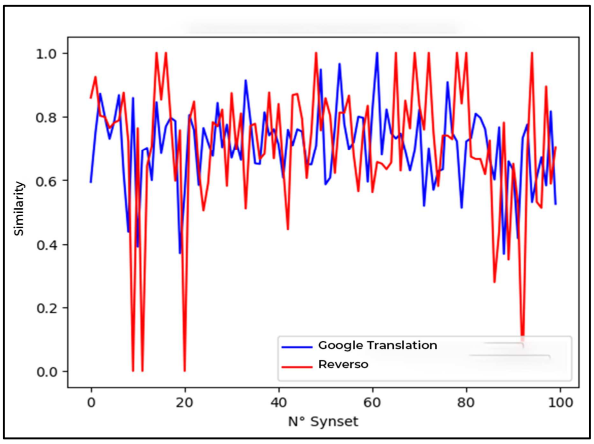


Figure 6: Comparison of translation similarities between Google Translate and Reverso

Analyzing the first graph, we concluded that Google Translate produces a much better translation quality compared to the MT5 model. Similarities are generally higher with Google Translate. On the other hand, the second graph indicates that Google Translate and Reverso provide translation quality that is almost similar, with only a few exceptions. This is why we have chosen to use Google Translate as the primary translator for the step of translating PWN definitions.

4.3 Definition filtering

The purpose of this step is to choose the best source of definitions between Wikidata and PWN. We used Google Translate to translate the definitions extracted from Wikidata into English. Then, we used BERT to calculate their similarities with the original definitions from PWN. The following graph highlights the similarities between the definitions from Wikidata, the translated definitions from PWN, and the original definitions from PWN.

The results show that the definitions extracted from Wikidata are not precise enough to capture the original context of the Synset. In fact, the similarities of the Wikidata definitions were at most 65%, whereas the translated definitions had similarities ranging from 65% to 100%, for the vast majority of these definitions. We thus chose to use the translated definitions for integration into the enriched AWN.

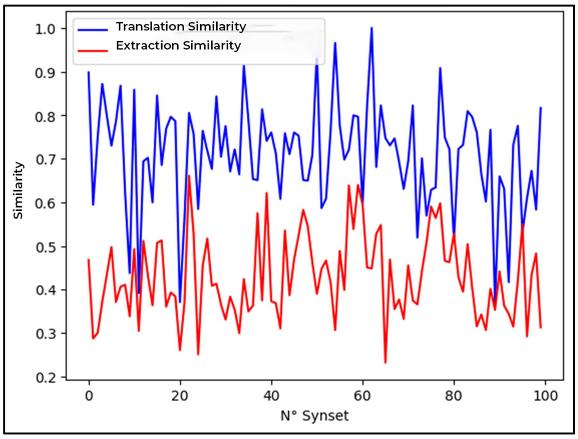


Figure 7: Similarity comparison between PWN and Wikidata

4.4 Example filtering

This step comprises two essential parts: (1) preprocessing the extracted data, and (2) selecting the examples to integrate into AWN.

Preprocessing data: As mentioned in the previous chapter, we used Reverso to extract the list of examples in Arabic (around 20 examples for each Synset translated from PWN). However, after manually analyzing some instances, we noticed the presence of several poorly formed and incorrect examples. To address this issue, we utilized the fine-tuned AraBERT model that we trained to classify sentences as either correct or incorrect. The results obtained were extremely satisfactory, demonstrating an outstanding performance of the classification model. However, despite achieving a high accuracy of 98% during model training, we observed that the lack of data led to a few misclassifications. Nevertheless, these errors did not pose a major obstacle to the next step. With the extraction of multiple examples for each translated word, we managed to compensate for any potential confusion. This approach allowed us to improve the quality of the examples and significantly reduce the size of the selected candidate examples, thus paving the way for more accurate and reliable results.

Selecting examples: After performing an initial filtering of the examples, we proceeded to the second step, which involves selecting the best examples, those that maximize contextual similarity with the senses of the Synsets translated into Arabic. For this purpose, we leveraged the AraBert model, which enabled us to calculate the similarity between the filtered examples and the definitions from PWN translated into Arabic using Google Translate. Once this step was completed, we selected, for each Synset, the most relevant example among all candidates. Table 2 represents some examples and their similarities with the definitions:

Synset ID	Definition in Arabic	Examples in Arabic	Similarity
able.a.01	يتبعها عادةً ("إلى") امتلاك الوسائل أو المهارة أو الدراية أو السلطة اللازمة للقيام بشيء ما	يعتقد البعض أن البنوك المركزية هي وحدها القادرة على ذلك	0.79
		من الناحية المثالية، سيكون لديك شريك قادر للمساعدة في التخطيط والتنفيذ	0.82
unable.a.01	يتبعها عادةً ("إلى") لا تمتلك الوسائل أو المهارة أو المعرفة اللازمة	وقال إن حكومة إسلام أباد تبدو وكأنها عاجزة أو رافضة للسيطرة على أراضيها	0.75
		وعندما تركت الكرسي المتحرك لاحظت أنها لا تستطيع استخدام الجانب الأيسر من جسدها	0.76

Table 2: Examples and their similarities with the definition

4.5 Word filtering

Similar to the selection of examples, we followed the same steps to choose the most appropriate translations for the sense of each Synset. Table 3 represents some words and their similarities with the definitions.

Synset ID	Definition in Arabic	Words in Arabic	Similarity
able.a.01	يتبعها عادةً ("إلى") امتلاك الوسائل أو المهارة أو الدراية أو السلطة اللازمة للقيام بشيء ما	إمكانية	0.47
		قادر	0.16
		جدير	0.15
unable.a.01	يتبعها عادةً ("إلى") لا تمتلك الوسائل أو المهارة أو المعرفة اللازمة	غير قادر	0.56
		لا تستطيع	0.52
abaxial.a.01	تواجه بعيدًا عن محور العضو أو الكائن الحي	محور	0.22

Table 3: Some words and their similarities with the definition

4.6 Construction of a Candidate Synset

Once we generated the lists containing the candidate translations (words, examples, definitions), we proceeded to create the XML file that would group the 90,127 candidate Synsets ready to be integrated into our validation platform database. Table 4 illustrates the obtained results:

Number of candidate Synsets	90,127
Number of candidate words added	231,165
Number of candidate examples added.	297,190
Number of candidate definitions added.	90,127
Number of Synsets with examples.	78,498
Number of words considered on average for each Synset	3
Average number of examples considered for each Synset.	3
Number of definitions considered for each Synset.	1

Table 4: Statistics of candidate Synsets.

4.7 Evaluation of the Results

In this section, we will delve into the results of the validation process conducted on a subset of 1,000 Synsets. Obviously, the validation is intended to be carried out by linguistics experts (and it will be done in the near future). However, in order to have a first assessment of the potential quality of our approach, we took on the responsibility of the manual evaluation of the sample. We consulted both the Collins English dictionary and the Elmaany Arabic dictionary to cross-reference the original English Synsets with their translated counterparts. Our validation process was facilitated through a dedicated platform. For word validation, we meticulously checked the English words, their contexts, and compared them with the translated Arabic versions. When it came to definitions, the majority were straightforward, allowing for a direct comparison between the English and Arabic definitions. In the case of examples, we meticulously selected the Arabic examples corresponding to the context of each word and proceeded with their validation.

Out of the randomly selected sample of 1000 Synsets, we obtained the following results:

	Validated number	**Rejected number**	**Precision**
Words	789	211	**78.9 %**
Definitions	952	48	**95.2 %**
Examples	813	187	**81.3 %**

Table 5: Precision of validation

5 Conclusion

In this work, we have presented a substantial enrichment of AWN along with a validation mechanism. Our contribution involved extending the currently available version of AWN on Global WordNet with automatically constructed Synsets based on PWN. We retrieved PWN Synsets, filtered out existing AWN equivalents, and translated them into Arabic, creating a list of candidate Synsets. Additionally, we extracted context-based examples for each Synset to enrich AWN. Deep learning techniques, particularly Transformers, were employed to evaluate and filter the Arabic Synset candidates. The results from the evaluation of a random sample of 1,000 Synsets taken from the candidate pool for enrichment are highly promising, with a manual validation accuracy of 75.4%. This work can further be enriched using other sources than just PWN, like Arabic text corpora. One could also consider using WordNets from other languages and look for ways of improving the quality of the translation.

References

Jorge Morato, Miguel Ángel Marzal, Juan Lloréns, and José Moreiro. 2004. In Proceedings of the Global WordNet Conf., GWC 2004, Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.), pages 270–278.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database, MIT Press. ed.

Horacio Rodríguez, David Farwell, Javi Farreres, Ma nuel Bertran, Musa Alkhalifa, Antonia Marti, William Black, Sabri Elkateb, James Kirk, Adam Pease, Piek Vossen, and Christiane Fellbaum. 2008. Arabic WordNet: Current State and Future Extensions. In Proceedings of the 4h Global WordNet Conf., Szeged, Hungary, pages 387–405.

William Black, Sabri Elkateb, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, Manuel Bertran, and Christiane Fellbaum. 2006. The Arabic WordNet Project. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06).

Musa Alkhalifa, and Horacio Rodríguez. 2018. Automatically Extending Named Entities coverage of Arabic WordNet using Wikipedia. International Journal on Information and Communication Technologies, 3(3): 20-36.

Mohamed Mahdi Boudabous, Nouha Chaâben Kammoun, Nacef Khedher, Lamia Hadrich Belguith, and Fatiha Sadat. 2013. Arabic wordnet semantic relations enrichment through morpho-lexical patterns. In Proceedings of the first International Conf. on Communications, Signal Processing, and their Applications, pages 1–6.

Mohamed Seghir Hadj Ameur, Ahlem Chérifa Khadir, and Ahmed Guessoum. 2017. An Automatic Approach for WordNet Enrichment Applied to Arabic WordNet. In International Conf. on Arabic Language Processing, pages 3–18.

Mohamed Ali Batita, and Mounir Zrigui. 2018. The Enrichment of Arabic WordNet Antonym Relations. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2017. LNCS, 10761, Springer, Cham.

Khang Nhut Lam, Feras Al Tarouti and Jugal Kalita. 2014. Automatically Constructing WordNet Synsets. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 106–111, Baltimore, Maryland, USA, June 23-25 2014.

Feras Al Tarouti and Jugal Kalita. 2016. Enhancing Automatic Wordnet Construction Using Word Embeddings. In Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP, pages 30–34, San Diego, California. Association for Computational Linguistics.

Mohamed Ali Batita, and Mounir Zrigui. 2019. The Extended Arabic WordNet: a Case Study and an Evaluation Using a Word Sense Disambiguation System. In Proceedings of the 9th Global WordNet Conference.