



# Sentence Transformers and DistilBERT for Arabic Word Sense Induction

Rakia Saidi<sup>1</sup> <sup>a</sup> and Fethi Jarray <sup>1,2</sup> <sup>b</sup>

<sup>1</sup>*LIMTIC Laboratory, UTM University, Tunis, Tunisia*

<sup>2</sup>*Higher institute of Computer Science of Medenine, Gabes University, Medenine, Tunisia*

**Keywords:** Clustering, Word Embedding, Word Sense Induction, NLP, BERT, Arabic Language.

**Abstract:** Word sense induction (WSI) is a fundamental task in natural language processing (NLP) that consists in discovering the sense associated to each instance of a given target ambiguous word. In this paper, we propose a two-stage approach for solving Arabic WSI. In the first stage, we encode the input sentence into context representations using Transformer-based encoder such as BERT or DistilBERT. In the second stage, we apply clustering to the embedded corpus obtained in the first stage by using K-Means and Agglomerative Hierarchical Clustering (HAC). We evaluate our proposed method on the Arabic WSI summarization task. Experimental results show that our model achieves new state-of-the-art on both the Open Source Arabic Corpus (OSAC)(Saad and Ashour, 2010) and the SemEval arabic (2017).

## 1 INTRODUCTION

In natural language processing (NLP), Word Sense Disambiguation (WSD) and Word Senses Induction (WSI) are two close tasks that aim to determine the sense of an ambiguous word. Given a sense inventory for each word such as Wordnet, WSD is a supervised task that aims to assign a sense to every ambiguous word. Given a target word (e.g., “Bank”) and a set of sentences containing the target (e.g., “he cashed a check at the bank”, “he sat on the bank of the river”), WSD is a supervised task that aims to cluster the sentences according to their senses. Unlike the supervised WSD, do not need to know the label (sense) of each sentence, but the sentences inside a cluster should be close to each other in terms of lexical similarity and far apart from sentences in other clusters. An example is shown in the figure 1.

In this paper we are concerned by WSI task and we seek to partition sentences into groups based on their semantic similarity.

In this study, the transformers model more especially BERT embedding was explored for Arabic WSI. A two-stage approach were designed where first we encoded sentences by transformer-based encoder and second we applied clustering algorithms to partition sentences. To the best of our knowledge, this is the first Deep neural network based approach for Arabic WSI.


The rest of the paper is organized as follows: Section 2 presents the state of the art. Our approach is explained in Section 3. The experimental setup is presented in Section 4. Results and discussion are presented in Section 5. We conclude this paper with a summary of our contributions and discuss future extensions.


## 2 STATE OF THE ART

The problem of arabic WSI or unsupervised word sense have been studied using a few methods.

(Rogati et al., 2003) defined a stemming model based on statistical machine translation, Its only training resources were an English stemmer and a short (10K phrases) parallel corpus. After the training phase, parallel text is not required. By letting the stemmer adapt to a chosen domain or genre, monolingual, unannotated material can be used to further enhance the stemmer. Rogati et al. presented results for Arabic and mentioned that the method can be used for any language that needs affix removal.

To address this specific issue, (Pinto et al., 2007) describe a method that relies on clustering of a self-expanded version of the original dataset. Using pointwise mutual information, the self-expansion technique replaces each term in the original corpus with a set of co-related terms. Pinto et al. mentioned that this concept, which was evaluated for the English language, performs well for the Arabic language as well,

<sup>a</sup>  <https://orcid.org/0000-0003-0798-4834>

<sup>b</sup>  <https://orcid.org/0000-0003-2007-2645>

- **S: (n) bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- **S: (n) depository financial institution, bank, banking concern, banking company** (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- **S: (n) bank** (a long ridge or pile) *"a huge bank of earth"*
- **S: (n) bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- **S: (n) bank** (a supply or stock held in reserve for future use (especially in emergencies))

Figure 1: Five senses of the word bank in WordNet.

demonstrating its linguistic flexibility.

For the word embedding approaches, (Djaidri et al., 2018) used the word2vec models. They investigated CBOV and Skipgram. Then, using the Annoy indexer, which is quicker than the Gensim similarity function, the model enables the construction of an indexer based on the cosine similarity. The graph is clustered to produce the various word meanings. They collaborated with OSAC and Aracorpus, two distinct news corporations. For a sample of Arabic ambiguous words, they mentioned that they had good results for word sense induction and good word sense discrimination performance.

Our method fits into the approaches of word embedding the only difference is that we exploited the models to transform more precisely BERT due to the limits of word2vec which are:

- **Training:** The networks' training differs significantly. A straightforward single-layered neural network called Word2Vec is trained using the ngrams of each distinct word as training data. A sentence from the corpus is used to train BERT to predict a masked word and the following sentence.
- **Vectors:** Word2vec stores a single vector representation of a word, whereas BERT creates a vector for a word based on its placement in a phrase or a sentence.

Besides, these word embedding methods were applied in several WSD works such as (El-Razzaz et al., 2021; Saidi and Jarray, 2022; Al-Hajj and Jarray, 2021; Saidi et al., 2022b; Saidi et al., 2022a), so we want to know its results on the WSI since these two problems are very close.

### 3 PROPOSED APPROACH

We propose a two stage clustering approach for WSI. First, we generate the sentence embedding by fine tuning BERT and DistilBERT framework. Second, run clustering algorithms such as K-means and HAC sentence embedding vectors.

Our model using k-means algorithm is presented in figure 2.

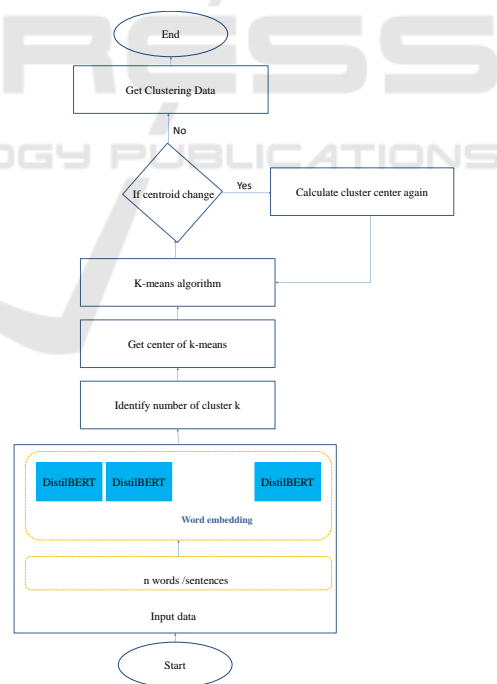


Figure 2: Arabic WSI model based on K-Means algorithm and DistilBERT word embedding.

### 3.1 Sentence Embedding

Word embedding refers to feature learning techniques in NLP where words are mapped to distributed dense vectors. Different word embedding techniques have been proposed in the literature such as word2vec, GLOVE, Elmo and BERT. Similarly, sentence embedding refers to a vector representation of an entire sentence. It can be obtained by the aggregation of its words embeddings or directly by creating a dummy vector that represents the sentence.

After reloading the data, we used DistilBERT from SentenceTransformers:

1. **BERT.** Bidirectional Encoder Representations from Transformers (BERT) is an unsupervised language representation. It has been successfully used in different NLP tasks, such as sentiment analysis (Chouikhi et al., 2021) and documents summarization (Tanfour and Jarray, 2022). Practically, we input a sentence into BERT and we get the vector representation of the sentence as the hidden representation of the special classification token ([CLS])
2. **DistilBERT.** DistilBERT (Sanh et al., 2019) is a general-purpose pre-trained version of BERT, 40% smaller, 60% faster, that retains 97% of the language understanding capabilities.
3. **Sentence-BERT (SBERT).** Sentence-BERT (SBERT) is an extension of the BERT model based on siamese network and triplet loss to generate semantically meaningful sentence embeddings. A Siamese Network is a deep learning network that contains two identical subnetworks used to generate feature vectors for each input and compute the similarity between the two inputs.

Figure 3 shows the difference between BERT model and DistilBERT model.

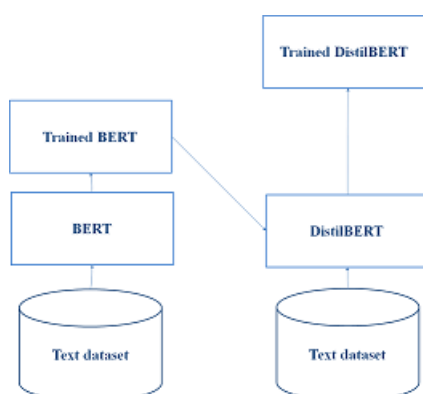


Figure 3: DistilBERT vs BERT.

### 3.2 Sentence Clustering

Sentence clustering consists in dividing a textual corpus into groups of semantically similar sentences. In WSI, ideally, each sense will be assigned to a different cluster. We cluster sentence embedding through K-Means and hierarchical agglomerative clustering HAC.

K-Means is an unsupervised Machine Learning algorithm that aims to partition data points into K clusters of equal variance. It alternates between the assignment of the data points to the nearest clusters while keeping the centroid of the clusters fixed, and updating the centroid centers while holding the assignments fixed. HAC consists in iteratively merging the two nearest pairs of clusters by the first step. The second step of our approach consists on applying K-Means and HAC to cluster the sentences embedding obtained by the first step. The main advantage of hierarchical clustering over K-Means clustering is that it is not necessary to prespecify the number of clusters and it can be applied to both categorical and numerical features. However, HAC may be slow for very large datasets due to the updates of the distance matrix at each iteration and it may be less efficient when clusters have a hyper spherical shape.

### 3.3 WSI Evaluation

This is the first Arabic work for WSI that uses a metric for evaluation. Clustering validation has been recognized as one of the important factors essential to the success of clustering algorithms. How to effectively and efficiently assess the clustering results of clustering algorithms is the key to the problem.

We used the internal cluster validation index Calinski-Harabasz (CH) Index. CH-index can be used to evaluate the model when ground truth labels are not known, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset. The CH-index also known as Variance Ratio Criterion (VRC) is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Here cohesion is estimated based on the distances from the data points in a cluster to its cluster centroid and separation is based on the distance of the cluster centroids from the global centroid.

The CH-index for K clusters on a dataset  $D = [d_1, d_2, d_3, \dots, d_N]$  is defined as,

$$CH = \left[ \frac{\sum_{k=1}^K n_k \|c_k - c\|^2}{K-1} \right] / \left[ \frac{\sum_{k=1}^K \sum_{i=1}^{n_k} \|d_i - c_k\|^2}{N-K} \right]$$

Table 1: CH-index for Arabic WSI model.

Dataset	#Cluster	CH on K-Means	CH on HAC-Ward	CH on HAC-Single	CH on HAC-Complete	CH on AHC-AVG
OSAC	5	238.83	192.35	7.29	11.563	11.61
	10	207.43	174.75	6.51	50.55	6.74
	50	85.03	78.32	2.78	58.14	52.08
SemEval	15	183.42	82.56	6.45	82.56	16.76
	34	105.49	61.39	7.68	61.39	12.67
	50	81.72	75.81	7.45	51.65	13.12

where,  $n_k$  and  $c_k$  are the number of points and centroid of the  $k$ th cluster respectively,  $c$  is the global centroid of the dataset,  $N$  is the total number of data points.

## 4 EXPERIMENTAL RESULTS AND EVALUATION

### 4.1 Dataset

We valid our approach on two Arabic datasets :The Open Source Arabic Corpus (OSAC)(Saad and Ashour, 2010) and the SemEval arabic task<sup>1</sup>.

- **OSAC.**

It is a corpus constructed from many websites. It is split into three primary categories: Following the elimination of stop words, the BBC-Arabic Corpus has 1,860,786 (1.8M) words and 106,733 unique words, whereas the CNN-Arabic Corpus contains 2,241,348 (2.2M) words and 144,460 unique words. After stopping words were removed, OSAC, which was gathered from several sources(Saad and Ashour, 2010) , contained roughly 18,183,511 (18M) words and 449,600 unique words. It is divided into 10 categories.

This corpus is used in (Djaidri et al., 2018) as a baseline.

- **SemEval.**

We used a new version of SemEval (2017), it is split into three subtask: Message Polarity Classification (Subtask A), Topic-Based Message Polarity Classification (Subtasks B-C) and Tweet quantification (Subtasks D-E), it contains 2,278 for training, 585 for validation and 1,518 for test. We investigated just the training data. This data set contains 34 classes. This corpus is used in (Pinto et al., 2007) as a baseline.

<sup>1</sup>[https://www.dropbox.com/s/i9tkaajuq1qbgjq/2017\\_Arabic\\_train\\_final.zip?](https://www.dropbox.com/s/i9tkaajuq1qbgjq/2017_Arabic_train_final.zip?)

### 4.2 Experimental Results

We run the following experiments to study the different aspects of the proposed approach. We automatically clean and cluster the datasets as the following:

- Experiment 1 : the number of cluster equal to the number of class existing in the dataset (10 for OSAC and 34 for Arabic SemEval).
- Experiment 2: the number of clusters is greater than the number of classes existing in the dataset.
- Experiment 3 : the number of clusters is less than the number of classes existing in the dataset.

For the HAC clustering algorithm, we adopt four similarity measures: ward link, single link, complete link and average link. For all experiments, we choose to do the same number of samples (5000), it helps us after to do a credible comparison between the obtained results.

Our main experimental results using CH-index metric are shown in Table 1. The CH-index on K-Means outperforms the CH-index on HAC and the ward linkage outperforms the other linkage types,

For SemEval data and when the number of cluster is 34, all the sentences containing the same word or the same context as this word are put into the same cluster, for example the sentences related to the word « أندرويد » are all put in the cluster 4.

Because the ward linkage performs better than others, we choose to plot some clusters points.

We tested also AHC algorithm with un-predefined number of cluster. Figure 5 plots the embedding data with Agglomerative Clustering with 3 clusters and with no predefined clusters.

The dendrogram (with complete(1) ,ward(2) single(3) and average(4) linkage) are presented in figure 6 for SemEval embedding data and in figure 7 for OSAC.

Table 1 shown that the CH-index on K-Means for OSAC performs better than on SemEval, so we can note that our system gives better results in word sense clustering (OSAC dataset) than in sentences

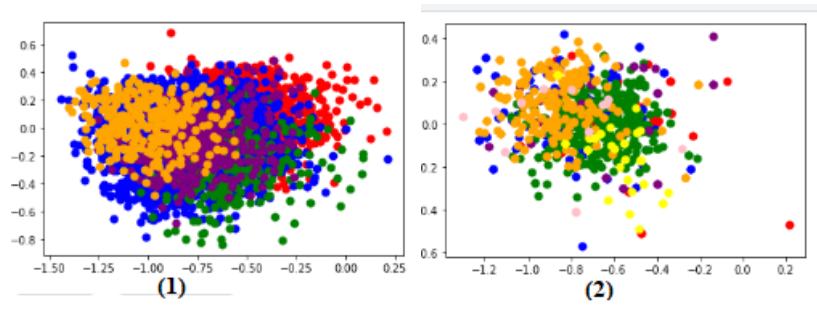


Figure 4: AHC for each embedding data, 10 for OSAC(1) and 34 for arabic SEMEVAL(2), this figure shows 5 clusters.

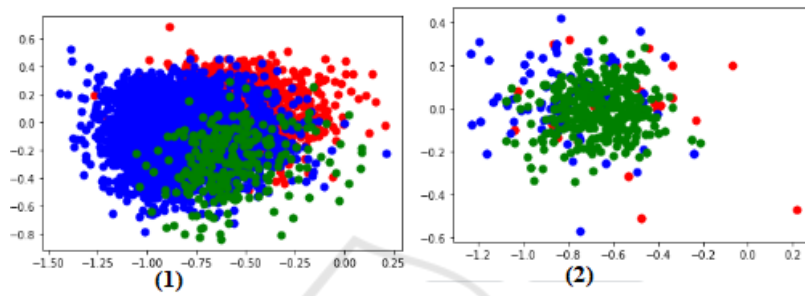


Figure 5: Result without predefined number of cluster, OSAC(1) and SEMEVAL(2).

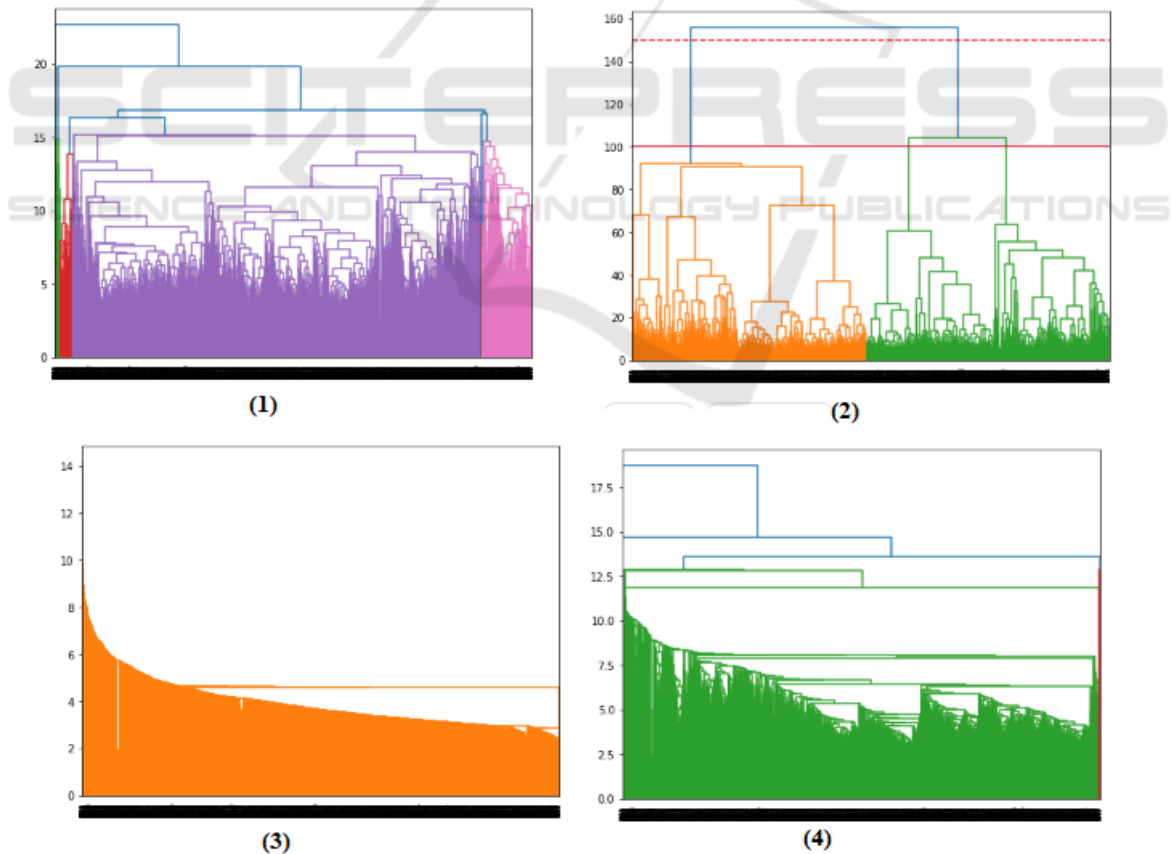


Figure 6: Dendrogram for SemEval data embedding.

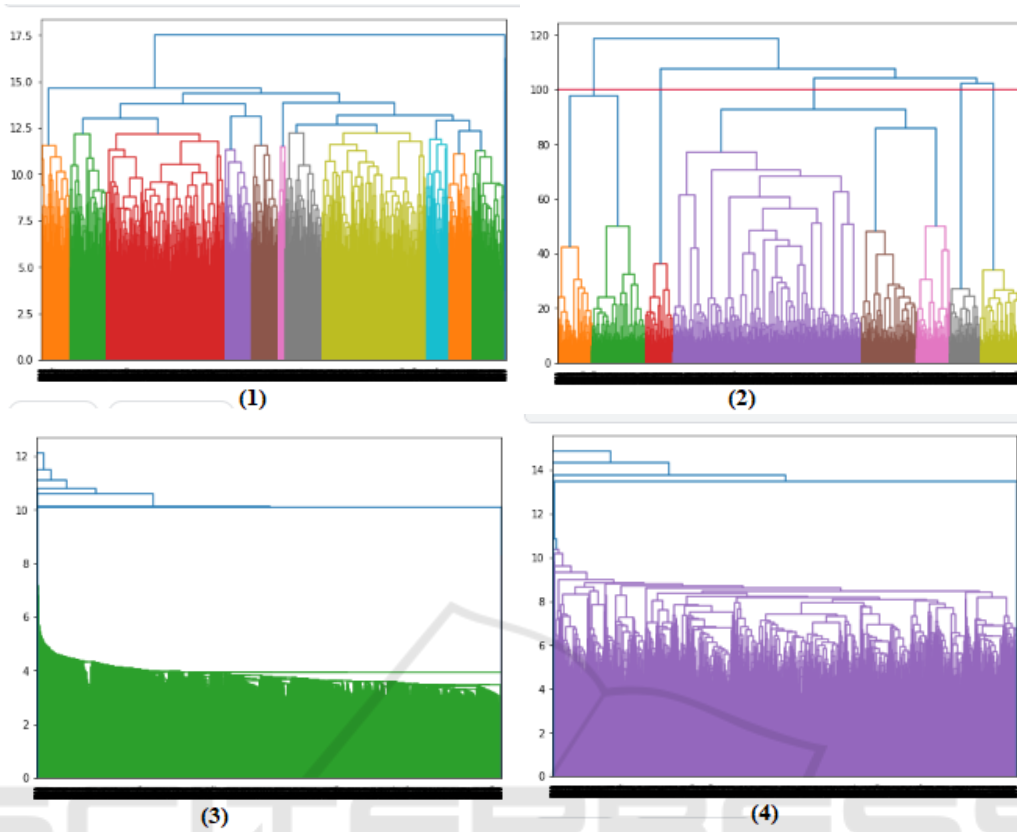


Figure 7: Dendrogram for OSAC data embedding.

Table 2: CH of an example by updating the number of samples.

#Example	#Cluster	CH
2000	3	284.94
	10	131.84
	50	57.49
	100	41.17
	150	34.97
3000	3	346.62
	10	177.79
	50	75.78
	100	54.35
	150	45.76
4000	3	<b>387.62</b>
	10	198.72
	50	93.54
	100	66.12
	150	55.07

(Semeval). Based on this result, we chose to test a set of words by modifying each time the number of examples as well as the number of clusters. These tests are shown in table 2.

If we have a higher number of examples the CH-index increases, if we have less number of cluster we

have a good CH-index.

#### Number of Examples: 4000

For example, the word «إثبات» and the word «إستجاب» belong in the same cluster when the number of cluster equals to 10, but each word belongs in a cluster when the number of clusters increases (150 here).

With 10 clusters:

in Cluster 3:

إنسكب،، «إثبات»، «إنسحاب»، «إنتخب»،  
 «إنتبه»، «إنتدب»، «إنتخب»، «إنتخب»، «إنتصاب»،  
 «إنتباه»، «إنتباه»، «إنتباه»، «إنتباه»، «إنتباه»،  
 «إنتداب»، «إنتخاب»، «إقتراب»، «إرتاب»، «إرتبط»،  
 «إرتكب»، «إرتباط»، «إرتباط»، «إرتباط»، «إرتباط»،  
 «إرتباط»، «إرتباط»، «إرتباط»، «إرتباط»، «إرتباط»

Table 3: Eye Clustering.

Cluster 1	Cluster 2	Cluster 3
العين هي مدينة تابعة لإمارة أبوظبي في الإمارات العربية المتحدة	مجموعة كلمات بحرف العين لتعليم الأطفال	تتألف عين الانسان من ثلاثة طبقات
مدينة العين هي مقر ممثل حاكم أبوظبي	كلمات بحرف العين بالصور للأطفال	العين هي المسؤولة عن الإبصار
تمثل مدينة العين رابع أكبر مدن الإمارات.	تدريب أقرأ لدرس حرف العين	العين عضو من أعضاء الجسم
تُوصف مدينة العين بكونها مدينة الحدائق في دولة الإمارات	أكتب كلمة تتضمن حرف العين	العين هي عضو يلتقط الضوء الذي تعكسه الأشياء

With 150 clusters:

In Cluster 1:

«أثبت»، «اثبات»

In the cluster 45:

«إستجابة»، «إستجلاّب»، «إستجواب»، «إستبدل»، «إستجاب»، «إستقلب»، «إستوعب»، «إستوجب»

We can observe and note that when the number of clusters is higher, the senses of words are closer and belong to the same cluster and for the sentences when the number of clusters is higher and the same word is used in different contexts, our system is able to eliminate the ambiguity of meaning and put each group of sentences with the same meaning together in a cluster and other sentences in another cluster despite all the sentences containing the same word. We take the word عين (eye in english) as an example. This word in Arabic has 100 meanings and each one can be distinguished according to the context containing عين. We used three senses in different sentences. Figure 8 present the set of sentences example and table 3 present its clustering.

We observed that the set of sentences was clustered into three clusters: cluster 1 eye is a city in Emirat, cluster 2 eye is an alphabet letter in the Arabic language and cluster 3 eye is the organ.

## 5 CONCLUSION

In this paper, we have proposed a novel two-step approach for Arabic word sense induction. First, we encode every sentence by transformer-based encoder. Second, we cluster the embedded sentences by clus-

العين هي مدينة تابعة لإمارة أبوظبي في الإمارات العربية المتحدة

مجموعة كلمات بحرف العين لتعليم الأطفال

تتألف عين الإنسان من ثلاثة طبقات

مدينة العين هي مقر ممثل حاكم أبوظبي

كلمات بحرف العين بالصور للأطفال

العين هي المسؤولة عن الإبصار

تمثل مدينة العين رابع أكبر مدن الإمارات

تدريب أقرأ لدرس حرف العين

العين عضو من أعضاء الجسم

تُوصف مدينة العين بكونها مدينة الحدائق في دولة الإمارات

أكتب كلمة تتضمن حرف العين

العين هي عضو يلتقط الضوء الذي تعكسه الأشياء

Figure 8: Eye sentences example.

tering algorithms such as k-means and HAC. We evaluate the model on both OSAC and Semval datasets. The experimental results achieve state-of-the-art performance through Calinski-Harabasz (CH) Index with 238.83 on K-Means algorithm and 192.35 on HAC-Ward linkage method for the OSAC dataset.

## REFERENCES

- Al-Hajj, M. and Jarrar, M. (2021). Arabglossbert: Fine-tuning bert on context-gloss pairs for wsd.
- Chouikhi, H., Chniter, H., and Jarray, F. (2021). Arabic sentiment analysis using bert model. In *International*

- Conference on Computational Collective Intelligence*, pages 621–632. Springer, Cham.
- Djaidri, A., Aliane, H., and Azzoune, H. (2018). A new arabic word embeddings model for word sense induction. In *19th International Conference on Computational Linguistics and intelligent Text Processing, CI-Cling*.
- El-Razzaz, M., Fakhr, M. W., and Maghraby, F. A. (2021). Arabic gloss wsd using bert. *Applied Sciences*, 11(6):2567.
- Pinto, D., Rosso, P., Benajiba, Y., Ahachad, A., and Jiménez-Salazar, H. (2007). Word sense induction in the arabic language: A self-term expansion based approach. In *Proc. 7th Conference on Language Engineering of the Egyptian Society of Language Engineering-ESOLE*, pages 235–245.
- Rogati, M., McCarley, J. S., and Yang, Y. (2003). Unsupervised learning of arabic stemming using a parallel corpus. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, pages 391–398.
- Saad, M. K. and Ashour, W. M. (2010). Osac: Open source arabic corpora. In *6th ArchEng Int. Symposiums, EEECS*, volume 10.
- Saidi, R. and Jarray, F. (2022). Combining bert representation and pos tagger for arabic word sense disambiguation. In *International Conference on Intelligent Systems Design and Applications*, pages 676–685. Springer.
- Saidi, R., Jarray, F., and Alsuhaibani, M. (2022a). Comparative analysis of recurrent neural network architectures for arabic word sense disambiguation. In *WE-BIST*, pages 272–277.
- Saidi, R., Jarray, F., Kang, J., and Schwab, D. (2022b). Gpt-2 contextual data augmentation for word sense disambiguation. In *PACIFIC ASIA CONFERENCE ON LANGUAGE, INFORMATION AND COMPUTATION*.
- Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Tanfouri, I. and Jarray, F. (2022). Genetic algorithm and latent semantic analysis based documents summarization technique. In *KDIR*, pages 223–227.