Network-based Approach for Stopwords Detection

Felermino D. M. A. Ali[1,3], Gabriel de Jesus[2]
Henrique Lopes Cardoso[2], Sérgio Nunes[2], Rui Sousa-Silva[3]
[1]Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
[2]Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC)
[3]Centro de Linguística da Universidade do Porto (CLUP)
{up202100778, hlc, sergio.nunes}@fe.up.pt
gabriel.jesus@inesctec.pt, rssilva@letras.up.pt

Abstract

Stopword lists, an essential resource for natural language processing and information retrieval, are often unavailable for low-resource languages. Creating these lists is time-consuming and expensive, making automated stopword detection a viable alternative. This paper introduces a novel stopword detection approach that exploits the topological properties of co-occurrence networks to identify function words. By leveraging the connectivity patterns of function words in these networks, the proposed approach aims to achieve higher precision compared to traditional frequency-based methods. To assess the effectiveness of the network-based approach, we constructed co-occurrence networks for Tetun and Emakhuwa (low-resourced languages), as well as English and Portuguese. We then compared the performance of this approach with traditional frequency-based methods. The results indicate that the network-based approach consistently outperforms traditional methods, with in-degree emerging as the most reliable indicator of function words. This finding suggests promising prospects for automatically generating stopword lists in other low-resource languages, paving the way for developing natural language processing tools for these linguistic contexts.

Keywords: Stopwords detection, Low-resource languages, Tetun, Emakhuwa.

1 Introduction

In natural language processing (NLP) and information retrieval (IR), stopwords are function words, such as prepositions, pronouns, and conjunctions, that are frequently removed due to their high frequency and minimal information content. A common practice in various NLP and IR tasks is to remove stopwords during the preprocessing stage to reduce execution time, enhance overall performance, and improve the effectiveness of retrieval systems (Croft et al., 2009).

Many existing methods often either rely on a predefined list of stopwords or are computed using traditional unsupervised methods, such as term frequency (TF) (Baeza-Yates and Ribeiro-Neto, 2011; Croft et al., 2009), normalized inverse document frequency (NIDF) (Lo et al., 2005), inverse document frequency (IDF), term frequency-inverse document frequency (TF-IDF), and term and document frequency (TDF) (Ferilli, 2021). Considering the observed high topological connectivity of function words in network properties (Chen et al., 2018; Gao et al., 2014; Liang et al., 2009), our objective is to investigate the application of co-occurrence networks' properties for automated stopword detection. This investigation is grounded on the assumption that the attributes of a co-occurrence network may prove more effective than traditional unsupervised frequency-based approaches in stopword detection tasks.

We employed co-occurrence network methods to validate the assumption, assuming two sequential terms in the text corpus form pairs of linking nodes and the graph is directed (see Figure 1). The datasets used for the experiment were collected from four languages: English, Portuguese, Tetun, and Emakhuwa. Tetun is one of Timor-Leste's official languages alongside Portuguese, spoken by 79.04% of a 1.17 million population (de Jesus, 2023), while Emakhuwa, also known as Makua, Macua, or Makhuwa, is a Bantu language primarily spoken in the northern and central areas of Mozambique with an estimated 7 million speakers (Ali et al., 2021).

Subsequently, these datasets underwent preprocessing to align with the requirements of our task. Following that, we constructed the directed co-occurrence network and evaluated the effectiveness of network attributes against traditional unsupervised frequency-based methods using precision at different levels. The results demonstrated that

our proposed approach outperformed all frequency-based methods for stopword detection in both high- and low-resource languages. This outcome implies the adaptability and applicability of our solution to automate the stopword list construction process in other low-resource languages, providing a valuable and efficient tool for language processing tasks in diverse linguistic contexts.

The remaining sections of this paper are organized as follows. Section 5 describes related works. The approach is outlined in Section 2. Then, Section 3 presents the experiment and evaluation. Section 4 presents the results obtained and their discussion. Finally, Section 6 summarizes our conclusion and possible future work.

2 Approach

We experimented with English, Portuguese, and two low-resourced languages (Emakhuwa and Tetun). We pre-processed the raw text data in both cases, then constructed a co-occurrence network by generating a directed graph $G = (V, A)$ from pre-processed text data. V denotes the nodes (i.e., word types) and A the edges. Each node $v \in V$ corresponds to a word from the vocabulary of the pre-processed text, whereas $a \in A$ is an adjacency arc, which represents the connection between a pair of consecutive words, from the first to the second.

Since we are interested in analyzing the network properties and unsupervised approaches concerning stopword detection, we compute the following node (i.e., word type) attributes:

1. **Network properties**: degree, indegree, outdegree, weighted indegree, weighted outdegree, closeness centrality, harmonic closeness centrality, eccentricity, and betweenness centrality.

2. **Traditional unsupervised methods**: term frequency (TF), normalized term frequency (NTF), inversed document frequency (IDF), document frequency (DF), normalized inverse document frequency (NIDF), term frequency - normalized inverse document frequency (TF-IDF), and term-document frequency (TDF).

For convention purposes, we use the following notation: C corresponds to the corpus of each language, and $n = |C|$ the number of sentences of corpus C, while $V = t_1, ..., t_m$ is the vocabulary of C, i.e., its word types (unique words) in C. For each term $t_i \in V$, o_i corresponds to the number of occurrences in C, n_i to the number of sentences in which it appears, and o_i^c to the number of occurrences in sentence $c \in C$. Thus, the total number of tokens in C is given by $o = \sum_i o_i$. In addition, from the network perspective, we denote M as an adjacency matrix, which contains boolean values indicating if there is a direct link between node i and j, for $m_{ij} \in M$.

The detail of network properties is explained in Section 2.1. Section 2.2 presents the traditional metrics. Finally, Sections 2.3 and 2.4 describe the data collection and preparation processes.

2.1 Network properties

The details of network properties are the following:

In-Degree The in-degree of node t_i is the total number of connections onto node i, and is the sum of the ith row of the adjacency matrix M:

$$t_i^{in} = \sum_j m_{ij} \tag{1}$$

Out-Degree The out-degree of node t_i is the total number of connections coming from node i, and is the sum of the ith column of the adjacency matrix M:

$$t_i^{out} = \sum_j m_{ij} \tag{2}$$

Degree The sum of all connections to the nodes t_i.

$$t_i^{degree} = t_i^{in} + t_i^{out} \tag{3}$$

Average Degree is simply the mean of all the node degrees in a network.

$$\frac{1}{n} \sum_{i=1}^{n} t_i^{degree} \tag{4}$$

Average Weighted Degree sum of the weights of all links attached to node i.

$$\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w(i,j)}{n} \tag{5}$$

Average Path Length the average length of shortest paths between any two nodes.

$$\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} d(i,j)}{n(n-1)} \tag{6}$$

Where d(i, j) is the shortest path length between nodes i and j, and n is the number of nodes in the network.

Original	Eusébio (25 january 1942 – 5 january 2014) was a Portuguese football player. He was born in Mozambique
Pre-processed	eusébio january january he was a portuguese football player he born in mozambique
Vocabulary	eusébio, january, he, was, a, portuguese, football, player, born, in, mozambique

Table 1: Text pre-processing example.

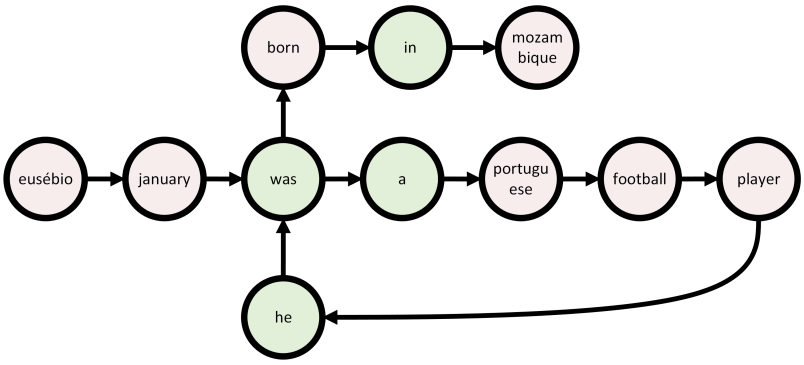

Figure 1: Co-occurrence network generated from the pre-processed text in Table 1.

Betweenness Centrality Measure the amount of influence that a node t_i has over the flow of information in the network:

$$C_B(i) = \sum_{j<k} g_{jk}(j) g_{jk} \tag{7}$$

When Normalized:

$$C_B(i) = \frac{\sum_{j<k} \frac{g_{jk}(j)}{g_{jk}}}{(N-1)(N-2)} \tag{8}$$

Where g_{jk} denotes the number of shortest paths connecting nodes j and k, $g_{jk}(i)$ is the number of those paths that pass through node i, and N is the number of nodes in the giant component.

2.2 Traditional unsupervised methods

The conventional frequency-based unsupervised approaches commonly used in stopword detection tasks are the following:

Term Frequency (TF) The amount of times a term appears in the corpus:

$$\mathrm{tf}(t_i) = o_i \tag{9}$$

Document Frequency (DF) The number of sentences in which a term occurs:

$$\mathrm{df}(t_i) = n_i \tag{10}$$

Normalized Term Frequency (NTF) TF normalized in accordance with the number of tokens in the corpus as a whole:

$$\mathrm{ntf}(t_i) = -log(\frac{o_i}{o}) \tag{11}$$

Inverse Document Frequency (IDF) (Church and Gale, 1995) Based on how frequently the term occurs in the corpus of sentences, the more it appears in sentences, the less information there is:

$$\mathrm{idf}(t_i) = log(\frac{n}{n_i}) \tag{12}$$

Normalized IDF (NIDF) (Robertson and Jones, 1976) IDF adjusted by 0.5 to reduce extreme values in relation to the number of sentences that do not contain the term $(n - ni)$:

$$\mathrm{nidf}(t_i) = log(\frac{(n - n_i) + 0.5}{n_i + 0.5}) \tag{13}$$

Term frequency-inverse document frequency (TF-IDF) (Sparck Jones, 1972) The product of NTF and NIDF:

$$\mathrm{TF} * \mathrm{IDF}(t_i) = \mathrm{ntf}(t_i) \times \mathrm{nidf}(t_i) \tag{14}$$

Term Document Frequency (TDF) (Ferilli, 2021) Amount of times a term appears in the corpus times the number of sentences in which it appears:

$$\mathrm{tdf}(t_i) = o_i \cdot n_i \tag{15}$$

2.3 Data Collection

We collected the datasets for Portuguese, English, Tetun, and Emakhuwa from different data sources. For Portuguese and English, we used WikiCLIR (Sasaki et al., 2018), a dataset containing Wikipedia articles from 25 different languages. It is one of the large-scale datasets constructed from Wikipedia's articles, making it ideal for our experiment as Wikipedia articles have wide coverage in terms of topics.

On the other hand, unlike English and Portuguese, datasets for Tetun and Emakhuwa are not easily accessible. That is why Tetun's dataset is entirely composed of news articles. These articles were extracted from two news online portals in Timor-Leste, Timor News[1]; and Tatoli[2]. We scraped the content using BeautifulSoup[3] and then built the corpus.

The Emakhuwa dataset is composed of a collection of the following data: Emakhuwa side of the parallel corpora from Ali et al. (2021) as well as Wikipedia articles[4], radio transcripts, and Emakhuwa's translation of the Constitution of the Republic of Mozambique. The details of the datasets are shown in Table 2.

2.4 Data Preparation

Since the datasets are collected from various data sources, we pre-processed them to exclude unnecessary characters and reduce the data size to be more efficient. The pre-processing task comprises lowercasing, tokenization, and removing punctuation, special characters, and numbers.

For Portuguese, English, and Emakhuwa, we used the general white space and punctuation-based tokenizer from NLTK[5] and a Tetun tokenizer[6] was used for Tetun.

Extra cleaning was done to reduce the vocabulary in Portuguese and English, as we noticed that Wikipedia articles typically contain a mixture of words from different languages. This happens because Wikipedia documents are usually translations of articles originally written in a different language, so some words, such as names and nouns, are kept as they are in the source languages. Thus, we removed all unknown terms from the language's word list to reduce the vocabulary. For Portuguese, we removed terms that do not appear in the Natura (University of Minho, 2021) dictionary, whereas for English, we used the english-words (Wiens, 2021) dictionary. The statistics for each network are displayed in Table 3.

3 Experiment and Evaluation

Our experiments consist of, first, conducting feature selection to reduce our analysis to the most relevant network attributes. After that, we subdivided the experiments into high-resource languages (English and Portuguese) and low-resource languages (Tetun and Emakhuwa). The experiments and evaluation are described in the following subsections.

3.1 Feature Selection

For each term in English and Portuguese vocabulary (i.e., nodes), we provided a Boolean value as true if they correspond to an actual stopword and false if they do not. For that, we used the stopword list available in the NLTK toolkit. Then, we load nodes and edges (i.e., co-occurrence) on Gephi[7] using the edge sum strategy and then computed the network's properties mentioned in Section 2.

To select the most relevant network features, we evaluated each network attribute with respect to information gain related to the target class (i.e., stopword or not stopword). Here, we use the Weka data mining toolkit (Frank et al., 2016) aiming to reduce our analysis to the top four relevant features. Weka calculates the information gain with respect to the target class by using the following formula:

$$InfoGain(c, a) = H(c) - H(c|a) \qquad (16)$$

Where $H(c, a)$ is the information for the dataset, c is the target class, a is the attribute, $H(c)$ is the entropy for the dataset before any change, and $H(c|a)$ is the conditional entropy for the dataset given the attribute a.

3.2 Portuguese and English

We run stopwords filtering based on the values computed from attributes in Section 2, following two strategies:

[1] https://www.timornews.tl

[2] https://tatoli.tl

[3] https://www.crummy.com/software/BeautifulSoup/

[4] https://incubator.wikimedia.org/wiki/Category:Wp/vmw

[5] https://www.nltk.org/

[6] https://pypi.org/project/tetun-tokenizer/

[7] https://gephi.org/

Language	#Documents	Data source(s)
English	50,000	WikiCLIR
Portuguese	50,000	WikiCLIR
Tetun	32,666	Timor News and Tatoli portals
Emakhuwa	52,238	Wikipedia, parallel corpora (Ali et al., 2021), radio transcripts, Mozambican's Constitution

Table 2: Dataset details. Documents are each line or paragraph of the corpus.

Measure	English	Portuguese	Tetun	Emakhuwa
# Nodes	18,369	57,960	22,111	53,880
# Edges	482,828	1,041,846	229,675	292,578
Diameter	7	11	21	17
Avg. Degree	26.285	17.976	10.387	5.43
Avg. Weighted Degree	150.934	84.277	49.534	14.215
Avg. Path Length	2.57	4.159	3.451	4.858

Table 3: Network statistics.

- Descending order (high to low): TF, NTF, DF, betweennesscentrality, indegree, outdegree, and degree.
- Ascending order (low to high): IDF, NIDF, and TF-IDF.

For evaluation, we adopt Precision N top position (P@N), which is given as the fraction of words that are stopwords:

$$P@N = \frac{stopwords}{N} \tag{17}$$

According to NLTK toolkit's stopword list, English has 126 stopwords, whereas Portuguese has 176. So, to evaluate precision, we used cutoff values from 25 ($P@25$) to 200 ($P@200$) with an interval of 25.

3.3 Tetun and Emakhuwa

Tetun and Emakhuwa do not have a ground truth list of stopwords, leading us to adopt the approach outlined in Section 3.2 for the stopwords filtering process. To assess precision, we strategically chose to translate the top 50 words into English, taking into account the widespread understanding and use of the English language. Following translation, each word was compared to entries in the English stopword list; if there was a match, it was classified as a stopword; otherwise, we considered it not to be a stopword.

4 Result and Discussion

We summarize our experimental results in Portuguese and English and discuss them in Section 4.1. In Section 4.2 is our observation of the approaches applied to Tetun and Emakhuwa.

Rank	Portuguese	Score
1	Betweeness Centrality	0.01223
2	Indegree	0.01162
3	Degree	0.01137
4	Outdegree	0.01123
5	TF-IDF	0.00988
6	Weighted Outdegree	0.00985
7	TF	0.00983
8	NTF	0.00983
9	IDF	0.00982
10	NIDF	0.00982
11	DF	0.00982
12	Weighted Degree	0.00979
13	Weighted Indegree	0.00973
14	TDF	0.00969
15	Harmonic Closeness Centrality	0.00753
16	Closeness Centrality	0.00730
17	Eccentricity	0.00128

Table 4: Feature ranking with information gain.

4.1 Portuguese and English

Table 5 and Table 4 provides the results of the importance of network features for stopword detection. For English, degree was the most relevant attribute, followed by in-degree. On the other hand, in Portuguese, betweenness centrality was at the top of the ranking. In-degree followed next, making a more stable feature as it ranked second in English and Portuguese. Then, degree and out-degree followed the list.

Based on these results, the betweenness centrality, indegree, outdegree, and degree are selected for the remaining experiments.

Table 6 shows results with the English network, where in-degree obtained the highest scores for all

Rank	English	Score
1	Degree	0.03525
2	Indegree	0.03515
3	DF	0.03342
4	NIDF	0.03330
5	IDF	0.03330
6	Outdegree	0.03298
7	TF	0.03272
8	Weighted Outdegree	0.03269
9	TF-IDF	0.03252
10	Weighted Degree	0.03238
11	NTF	0.03210
12	TDF	0.03170
13	Betweeness Centrality	0.03150
14	Weighted Indegree	0.03038
15	Harmonic Closeness Centrality	0.02912
16	Closeness Centrality	0.02795
17	Eccentricity	0.00632

Table 5: Feature ranking with information gain.

approaches. Similar results can be found on degree, the second best performing approach, outperforming all frequency-based approaches except $P@175$ and $P@200$. Betweness centrality, on the other hand, has provided results very close to frequency-based approaches; however, it has small margin advantages in $P@25$. Finally, out-degree was shown to be unworthy of English as it scored slightly below the frequency-based approaches.

Table 7 shows the results in Portuguese. Here, the network-based approaches have a clear advantage over frequency-based approaches. Betweenness centrality attained the highest scores, except for precision at the top 75 words. However, regarding complexity, betweenness centrality was the least efficient approach as it takes approximately $O(N^3)$ (Barthelemy, 2004) to compute the values. In-degree and degree, on the other hand, obtained similar results to betweenness centrality, where the degree was better than in-degree by small margins. This can be explained by the fact that degree is the summation of in-degree and out-degree, so the degree is always proportional to both in- and out-degree. However, like English, out-degree performed worse than all other network-based approaches.

Overall, our results support the claims by Gao et al. (2014); Chen et al. (2018); Liang et al. (2009) that stopwords are highly connected nodes in a co-occurrence network. However, we further suggest that degree and in-degree are more effective approaches for stopwords detection than unsupervised frequency-based ones.

4.2 Tetun and Emakhuwa

To investigate the precision of network-based properties for stopword detection further, we also experimented on Tetun and Emakhuwa. The results are presented in Table 8. Due to extra manual effort to project stopwords from the English language, we only focused on 50 top-ranking words. Here, the results show a clear advantage of network-based approaches over frequency-based ones. Also, in-degree outperforms the other approaches in Tetun and Emakhuwa. We visualized the top-25 stopwords for Tetun in Figure 2 and Emakhuwa in Figure 3. The network was filtered by in-degree and partitioned into stopwords (green) and not stopwords (red). Also, each node has a label that shows the original word and its translation, separated by a colon. The precision drop for the Emakhuwa language can be better visualized in Figure 3, which shows five misses. We believe that this drop in precision is because the Emakhuwa corpus is predominantly made up of religious texts, which contain a high number of words from religious themes, such as "Yesu" (meaning Jesus), "Muluku" / "Yehova" (meaning God), and others.

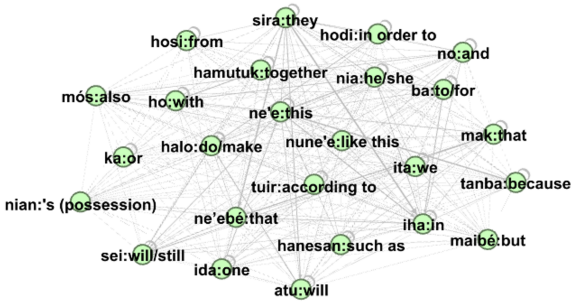

Figure 2: Top-25 of the Tetun stopwords identified using in-degree.

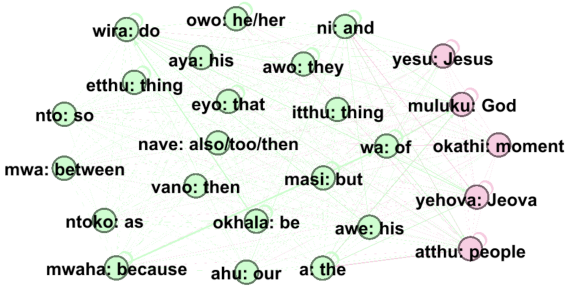

Figure 3: Top-25 stopwords of Emakhuwa identified using in-degree.

Approach	P@25	P@50	P@75	P@100	P@125	P@150	P@175	P@200
tf	0.960	0.800	0.706	0.600	0.528	0.480	0.451	**0.420**
ntf	0.960	0.800	0.706	0.600	0.528	0.480	0.451	**0.420**
df	0.920	0.800	0.706	0.620	0.544	0.500	0.463	0.410
idf	0.920	0.800	0.706	0.620	0.544	0.500	0.463	0.410
nidf	0.920	0.800	0.706	0.620	0.544	0.500	0.463	0.410
tf-idf	0.920	0.800	0.693	0.610	0.560	0.500	0.457	0.415
tdf	0.960	0.800	0.693	0.620	0.552	0.500	0.457	0.415
betweennesscentrality	**1.000**	0.800	0.693	0.620	0.552	0.487	0.429	0.390
indegree	**1.000**	**0.880**	**0.800**	0.690	**0.608**	**0.54**	**0.469**	0.415
outdegree	0.920	0.780	0.667	0.570	0.504	0.467	0.423	0.380
degree	**1.000**	0.840	**0.800**	**0.700**	**0.608**	0.520	0.460	0.411

Table 6: Stopword detection Precision for English.

Approach	P@25	P@50	P@75	P@100	P@125	P@150	P@175	P@200
tf	**1.000**	0.68	0.533	0.49	0.432	0.373	0.331	0.310
ntf	**1.000**	0.68	0.533	0.49	0.432	0.373	0.331	0.310
df	0.920	0.68	0.56	0.520	0.448	0.386	0.337	0.310
idf	0.920	0.68	0.56	0.520	0.448	0.386	0.337	0.310
nidf	0.920	0.68	0.56	0.520	0.448	0.386	0.337	0.310
tf-idf	0.920	0.68	0.546	0.520	0.448	0.373	0.337	0.315
tdf	0.960	0.680	0.546	0.510	0.440	0.373	0.331	0.310
betweennesscentrality	**1.000**	**0.920**	0.747	**0.620**	**0.552**	**0.480**	**0.440**	**0.400**
indegree	**1.000**	0.860	0.733	0.600	0.528	0.460	0.417	0.375
outdegree	0.960	0.800	0.707	0.590	0.512	0.447	0.411	0.370
degree	**1.000**	0.900	**0.773**	0.620	0.528	0.467	0.406	0.380

Table 7: Stopword detection Precision for Portuguese.

	Tetun		**Emakhuwa**	
Approach	P@25	P@50	P@25	P@50
tf	0.840	0.720	0.760	0.740
ntf	0.840	0.720	0.760	0.740
idf	0.880	0.760	**0.800**	0.760
nidf	0.640	0.500	**0.800**	0.760
tf-idf	0.880	0.760	**0.800**	0.760
tdf	0.880	0.740	0.760	0.760
betweennesscentrality	0.920	0.820	0.720	0.780
indegree	**1.000**	**0.900**	0.760	**0.820**
outdegree	0.960	0.840	0.760	0.800
degree	**1.000**	0.860	0.760	0.800

Table 8: Stopword detection Precision for Tetun and Emakhuwa

5 Related Works

The concept of stopwords was initially introduced by Luhn (1957) and as their application in the domains of information retrieval and natural language processing became evident, lists of stopwords were compiled for various languages. Several automated approaches for stopword detection have been proposed to streamline the process and eliminate manual effort. The conventional method for identifying stopwords uses term frequency (TF) (Manning et al., 2009; Croft et al., 2009). Lo et al. (2005) introduced the normalized inverse document frequency (NIDF) in their experiment with TREC[8]. These two techniques, along with other unsupervised frequency-based approaches, such as inverse document frequency (IDF) and term frequency-inverse document frequency (TF-IDF), among others, have served as the standard mechanism for stopword detection. More recently, Ferilli (2021) proposed another approach called term-document frequency (TDF), which proves particularly effective when dealing with small-sized corpora.

From a linguistic perspective, stopwords are function words, which serve grammatical or structural roles in sentences. These function words are typically high-frequency terms such as articles, prepositions, conjunctions, and pronouns, frequently occurring in conjunction with other words. This high-frequency terms has been validated across different languages through various studies (Chen et al., 2018; Gao et al., 2014; Liang et al., 2009), employing co-occurrence networks constructed based on the linear relation of words. Gao et al. (2014) in their analysis of six languages (Arabic, Chinese, English, French, Russian, and Spanish), observed that function words, including

[8] https://trec.nist.gov/

"y," "en," and "a," in the Spanish network, consistently ranked highest in degrees across all languages. Focusing on Chinese and English, Liang et al. (2009) reported that words with the highest connections (degree) were functional words, such as "a," "the," and "of," in English networks. Similarly, Chen et al. (2018) identified stopwords with the highest degree in Chinese co-occurrence networks, suggesting their role as hubs and indicating high betweenness centrality.

While the aforementioned studies offer evidence of a potential correlation between network properties and "stopwordness", to our knowledge, there has been no systematic evaluation of the advantages of complex network features in stopword detection. This is why this study investigates complex network properties for stopword detection.

6 Conclusion and Future Work

This paper presents an automated approach for stopword detection, leveraging network properties derived from connecting pairs of sequential terms within text corpora. The datasets underwent preprocessing before constructing pairs of sequential terms for selecting network features. Selecting these features involved evaluating the information gained from each one. The chosen network features were then employed in experimentation with the aforementioned four languages. Overall results indicate that network features are more effective than existing frequency-based approaches in stopword detection, with in-degree as the most reliable feature.

In future work, we aim to apply the proposed approach to construct stopword lists for Tetun and Emakhuwa. Furthermore, we will develop and use ground truth stopwords for both languages to conduct a more comprehensive evaluation of the effectiveness of the proposed approach.

Acknowledgements

This work was financially supported by Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC). Felermino Ali is supported by a PhD studentship (with reference SFRH/BD/151435/2021), funded by Fundação para a Ciência e a Tecnologia (FCT). Similarly, Gabriel de Jesus also benefits from a scholarship funded by FCT, identified by reference number SFRH/BD/151437/2021.

References

Felermino D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. 2021. Towards a parallel corpus of portuguese and the bantu language emakhuwa of mozambique. In *2nd AfricaNLP Workshop Proceedings, AfricaNLP@EACL 2021, Virtual Event, April 19, 2021*.

Ricardo Baeza-Yates and Berthier A. Ribeiro-Neto. 2011. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

Marc Barthelemy. 2004. Betweenness centrality in large complex networks. *The European Physical Journal B - Condensed Matter*, 38(2):163–168.

Heng Chen, Xinying Chen, and Haitao Liu. 2018. How does language change as a lexical network? an investigation based on written chinese word co-occurrence networks. *PLOS ONE*, 13(2):1–22.

Kenneth Church and William Gale. 1995. Inverse document frequency (IDF): A measure of deviations from Poisson. In *Third Workshop on Very Large Corpora*.

W. Bruce Croft, Donald Metzler, and Trevor Strohman. 2009. *Search Engines - Information Retrieval in Practice*. Pearson Education.

Gabriel de Jesus. 2023. Text information retrieval in Tetun. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III*, volume 13982 of *Lecture Notes in Computer Science*, pages 429–435. Springer.

Stefano Ferilli. 2021. Automatic multilingual stopwords identification from very small corpora. *Electronics*, 10(17).

Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. The weka workbench. online appendix for "data mining: Practical machine learning tools and techniques". Fourth Edition.

Yuyang Gao, Wei Liang, Yuming Shi, and Qiuling Huang. 2014. Comparison of directed and weighted co-occurrence networks of six languages. *Physica A: Statistical Mechanics and its Applications*, 393:579–589.

Wei Liang, Yuming Shi, Chi K. Tse, Jing Liu, Yanli Wang, and Xunqiang Cui. 2009. Comparison of co-occurrence networks of the chinese and english languages. *Physica A: Statistical Mechanics and its Applications*, 388(23):4901–4909.

Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. 2005. Automatically building a stopword list for an information retrieval system. *J. Digit. Inf. Manag.*, 3(1):3–8.

Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4):309–317.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Online edition, Cambrige UP.

Stephen E. Robertson and Karen Spärck Jones. 1976. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.*, 27(3):129–146.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463, New Orleans, Louisiana. Association for Computational Linguistics.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval".

UMINHO University of Minho. 2021. words-pt: Dicionário natura. Accessed: 2024-01-05.

Matt Wiens. 2021. english-words-py. Accessed: 2024-01-05.