Word Sense Disambiguation Pipeline Framework for Low Resourced Morphologically Rich Languages

Mosima Anna Masethe[1], *Hlaudi Daniel Masethe[2] , Sunday Olusegun Ojo[3] , and Pius A Owolawi[4]

[1]Sefako Makgatho Health Sciences University, South Africa

[2]Tshwane University of Technology, South Africa

[3]Durban University of Technology, South Africa

[4]Tshwane University of Technology, South Africa

Abstract

Resolving ambiguity problem is a prolonged natural language processing theoretical research challenge. Sesotho sa Leboa language is an official name for Sepedi or Northern Sotho language as known to be an official language among 11 others in South Africa spoken by 4.7 million people. Sesotho sa Leboa is an indigenous rich morphologically low resourced South African language which is a highly polysemous language, with words that have numerous context. Disambiguating polysemous words remain a challenging problem for computational linguistics research. Deficiencies of several polysemy assessments suggest that dealing with the sense distinctiveness versus polysemy problems remains an uncluttered academic issue. A practical problem in natural language processing applications is Word Sense Disambiguation which suffers drastically from shortcomings when working with ambiguous polysemous words. Therefore, Word Sense Disambiguation seeks both academic and practical results. Many Word Sense Disambiguation applications gives high accuracy for the English language, and poor accuracy for Sesotho sa Leboa language. In this research, Word Sense Disambiguation pipeline framework is developed for Sesotho sa Leboa low resourced morphologically rich language which addresses academic and practical problems of the polysemy problem. The proposed Word Sense Disambiguation pipeline framework shows pre-processing modules which is a process to reduce ambiguity from the unstructured text corpus that serve to input sentences. Hence, the researchers compute the probability of Word Sense Disambiguation when polysemy and homonymy is observed for cosine similarity measures using sentence transformer (SBERT) and Word2Vec algorithms (Skip-Gram and Continuous Bag of Words). Computation of cosine similarity measure shows SBERT outperforms other algorithms with 87% threshold which shows strong similarity between context and sense definition while Continuous Bag of Words gives cosine similarity threshold of 51%, outperforming Skip-Gram algorithms which has a threshold below 50% with two vectors approaching a perpendicular angle of 90-degrees orthogonally indicating that orientation of vectors do not match.

Keywords: Corpus, Continuous Bag of Words, Natural Language Processing, SBERT, Skip-Gram, Word Sense Disambiguation

1. Introduction

Word Sense Disambiguation (WSD) is a commission of categorising correct sense of the word used in a sentence, if the word has numerous senses (Baˇsi´c & ˇSnajder, 2018). Furthermore, WSD addresses both polysemy and homonymy words. It is also referred to as a method to find precise word similarity of an ambiguous word in a specific context and WSD methodologies are categorized into three main classes – Knowledge base, Supervised and Unsupervised approaches (Ranjan Pal & Saha, 2015).

Word sense ambiguity problem are caused by word with numerous senses referred to as polysemous words. Polysemous words are difficult to investigate in searching for the correct lexical category and the correct sense of the word. Kulkarni et al., (2012) argues that researchers in natural language processing (NLP) has recommended a pipeline technique to solve ambiguity problems. NLP pipelines can be processed separately, however, the researchers Jaafar & Bouzoubaa, (2016) in Arabic NLP recommends that one can call two or more pipeline tools in a sequential way, i.e. output of one pipeline can be an input into another pipeline tool. NLP pipeline components must be chosen based on a use-case sentence composition, which suggest that pipelines should be constructed with different strategies such as syntactic and semantic text analysis (Kulkarni et al., 2012).

WSD is identified to be a significant module in NLP pipeline which increase threshold of the obtained information (Torii et al., 2015). Agirre & Edmonds (2008) expresses WSD in NLP as the challenge of verifying which senses of the word is triggered by utilization of the word in a context, a practice active intuitively in human beings (Agirre & Edmonds, 2008). Nevertheless, devices process unstructured dataset and transform them into labelled data which must be evaluated in order to determine the correct senses (Popov, 2018). System that plans to manage natural languages as human beings do, should have context about words and their senses because senseful dataset are comprised of senseful words (Miller, 1995).

Sesotho sa Leboa language has 30 distinguishable dialects with conjunctive writing different to Nguni language, a linguistic word can have three orthographic units while Nguni linguistic word has bound and free morphemes which can be written as one linguistic word (FaaB, 2010).

Sesotho sa Leboa language is polysemous giving words that have various senses. Polysemy in Sesotho sa Leboa is of several types, including Part-of-Speech (POS), specialization, metaphoric, etc., each presenting challenges in WSD provisioning. Polysemy continues as a computing issue in public domain and linguistics research (Mmaseroka, 2006). As deficiencies of most polysemy assessments indicate that working with distinct senses and polysemy problem still is an open academic problem; and so, polysemy remains an open theoretical issue (Popov, 2018).

This paper outline proceeds first with introduction of WSD in relation to Sesotho sa Leboa polysemous words and orthographic writing. The second phase discuss literature review related to solutions that exists on WSD, how they have been applied and the existing approaches. Discussion on the research methodology follows with description of the experiment performed and presentation of the research result. Lastly the discussion and conclusions.

2. Related Literature Review

Word Sense Disambiguation (WSD) solutions has been provided through research for the English language, academic research has not much been done in Sesotho sa Leboa language, and many other African languages regarding word sense disambiguation. WSD solution is a key to designing Human language Technologies (HLTs) and Natural Language Processing systems. Sesotho sa Leboa language is considered a morphologically low resourced language, which is not well represented through digital language divide. It is imperative to leverage the language out through WSD solutions.

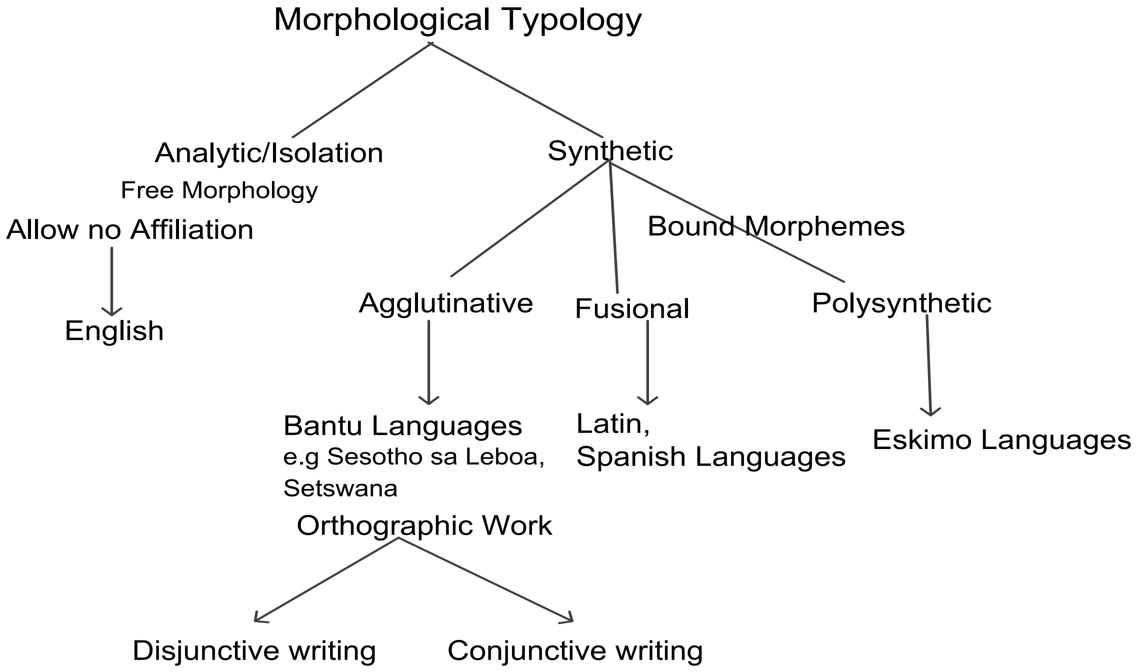
A number of WSD research focused on English language, and only few academic publications on WSD for Sesotho sa Leboa. A number of NLP pipeline tools like word segmentation, Part-of Speech tagging, Lemmatization, Multiword Detection and Named Entity Recognition has now been designed for Sesotho sa Leboa language, Nevertheless, the language lacks WSD pipeline to solve ambiguity. The NLP system is assessed in several levels namely: Morphological level, lexical level, syntactic level, semantic level, discourse level, and pragmatic level. The research considers WSD issues at the semantic level which focus on how the context of the words within a corpus help to define the sense of words (Açıkgöz et al., 2017)

South African languages are mostly considered morphologically low resourced with agglutinative data that can be used to design natural language processing (NLP) systems; the existing government has for the last 20 years gave assisted human language technologies (HLT) to develop NLP pipelines, which produced valuable language resources and applications. Language Resource Management Agency of the South African Centre for Digital Language Resources (SADiLaR) is tasked with language development in South Africa (Puttkammer et al., 2018).

Research by (Puttkammer et al., 2018) from the North-West University (NWU) designed NLP web services for ten South African languages which can be accessed through application programming interface (API) with a good web application. The technologies were designed and funded by National Centre for Human Language Technology (NCHLT) projects of South Africa in the Department of Arts and Culture.

Linguistic typology categorize language based on similarities in structure such as phonology, grammatical construction and word order, morphological process include affixation, to mention the few for creating words and word forms. Morphology is classified as a field of linguistic that study word structure and formation (Flanagan, 2013). Fig. 1, shows morphological typology (Giunchiglia, et al. 2018), which also classify English as an Analytic or Isolating theme with free morphemes which appear as independent words (Flanagan, 2013) . Further, Bantu language is classified as Agglutinative with bound morphemes which do not constitute independent words, but are attached to other morphemes or words. Bound morpheme are called affixes classified into – inflectional, derivational, prefixes, suffixes, infixes (Pirkola, 2001).

Figure.1 Morphological Typology



Sesotho sa Leboa, is a Bantu language in the Sotho group. The language is characterized by disjunctive orthography, mainly verb prefixal morphemes category. Suffixal morpheme follow conjunctive writing style, hence the language is considered to be semi-conjunctive. Furthermore, Bantu language is considered to be agglutinative and exhibit substantial inherent structural resemblance, though differ substantially in terms of orthography, due to both phonology and history (Pretorius & Pretorius, 2009; Flanagan, 2013).

The verb in Sesotho sa Leboa consists of infinitive prefix + a root + verb-final suffix. Example:

- Go bona (To see) – infinite prefix- go + root –bon, + and verb-final suffix –a
- Ba di bona (They see it) -
- Lesogana la sega (The young man then laughed) –
- Mošemane yo a kitimago (The boy who is running) – Infinitive prefix -Mo + root –šemane + root –Kitim + final suffix – a -go

Root is defined to be a lexical morpheme which is that part of a word that do not include grammatical morpheme and cannot occur independently as a word (Pretorius & Pretorius, 2009). In English structural words include prepositions, articles and pronouns generally assist verbs and nouns, whereas, in Sesotho sa Leboa many disjunctively written bound morphemes are function words and contribute to grammatical correctness of sentences (Rahab & Mothapo, 2019). The researchers cannot easily adopt a WSD tool for English to Sesotho sa Leboa as the language present new challenges compared to other Indo-European languages when defining the feature set.

WSD is a classification problem, where an ambiguous word is classified to its senses (Zopon et al., 2015). Target-word and All-word are approaches for WSD that disambiguate words using supervised approaches and unsupervised approaches (Zopon et al., 2015). The research aims to solve All-word WSD utilizing unsupervised classification methods.

The research consider polysemous words in which a lexical semantic contain more senses of the same word. For example, the word "Noka" in Sesotho sa Leboa has many senses. It can mean a hip or river or seasoning as applied below:

Mosadi o noka seshebo ka letswai (A woman is seasoning the food with salt)

Noka ya mma e bohloko (My mother's hip is painful)

Noka e tletse ka meetsi (The river is full of water)

Kulkarni et al. (2012) argues that researchers in NLP has adopted a pipeline approach to solve ambiguity problems. NLP pipelines can be processed separately, however, the researchers(Jaafar & Bouzoubaa, 2016) in Arabic NLP recommends that one can call two or more pipeline tools in a sequential way, i.e. output of one pipeline can be an input into another pipeline tool. NLP pipeline components must be chosen based on a use-case sentence composition, which suggest that pipelines should be constructed with different strategies such as syntactic and semantic text analysis (Kulkarni et al., 2012). NLP text-based application comprises of a pipeline of pre-processing steps such as tokenization, stemming, part-of-speech tagging, named entity recognition, chunking, parsing, the morphological analyser, phrasal marker, word sense disambiguator (WSD), scanner and sentence boundary identification, just to mention the few (Bangalore, 2006).

Supervised learning approaches process structured information in the arrangement of annotated training corpus (Pal & Saha, 2015). This technique use machine-learning algorithms from sense-annotated data created manually or semantically annotated corpora to introduce induction principle for classification models to determining the appropriate sense for each specific context (Pal & Saha, 2015; Navigli & Velardi, 2005).

Assume W = {w1, w2, … wn} as a data point with n features a set of All-word, the research aims to compute possible sense Sk to solve W among a set of {s1, s2, …, sk} senses. Eq. 1 defines Bayes theorem as:

$$P(S_k|W) = \frac{P(W|S_k)P(S_k)}{P(W)} \text{ for k =1, 2, . . .K} \quad (1)$$

Where P(Sk|W) is posterior probability and P(Sk) is prior probability of class, and P(W) prior probability of predictor. Applying Bayesian classifier to compute the correct senses is given by Eq. 2:

$$P(S_k) \prod_{i=1}^{n} P(wi|S_k)\, for\ k = (1, 2, \ldots, K) \quad (2)$$

Where P(wi|S_k) is the feature (Fj) in the context given by P(F1, …, Fj) assigning W to the sense Ś for the largest value. Expressed mathematically in Eq. 3:

$$\acute{S} = argmax_{k\in(1,2,\ldots,K)} P(S_k) \prod_{i=1}^{n} P(F_j|S_k) \quad (3)$$

There exist many approaches to solve WSD challenges for free morphology languages. Even though all approaches are excellent to solve the WSD problem, research study needs to prove that Eq. 3 can solve WSD problem for morphologically low resourced languages.

Skip-Gram model predict the context C_t around a given ambiguous word w_t (Sutor et al., 2019) expressed as:

$$C_t = w_{t-n}, \dots, w_{t-1,} \, w_{t+1}, \qquad \dots, w_{t+n}$$

Skip-Gram maximizes the total log probabilities across all T words expressed mathematically as:

$$\sum_{t=1}^{T} log\,(P(C_t|w_t))$$

Which is further condensed as:

$$log\,P(C_t|w_t)) = \sum_{j=-n}^{n} log\,(P\big(w_{t+j}\big|w_t\big)) \qquad (4)$$

The mathematical formula in Eq. 4 reflects a soft-max procedure and it’s asymptotically relational to the number of words.

The NB algorithm is a probabilistic model that makes use of Bayes rule and adopts conditional independence of features given the class label, it has been used considerable with success for WSD task (Bakx, 2006; Aliwy & Taher, 2019). NB approach categorizes text documents using two constraints named conditional probability of each sense (Si) of a word (w) and the features (fj) in the context (Bangalore, 2006; Shallu & Gupta, 2013).The appropriate sense in the context is represented by the maximum value assessed through the expressed formula (Pal & Saha, 2015).

$$\acute{S} = \begin{matrix} argmax \\ S_i \in Senses_D\,(w) \end{matrix} \; P\,(S_i|f_i, \dots, f_m)$$

$$\acute{S} = \begin{matrix} argmax \\ S_i \in Senses_D\,(w) \end{matrix} \; \frac{P(f_i, \dots, f_m|S_i)P(S_i)}{P(\,f_i, \dots, f_m)}$$

$$\acute{S} = \begin{matrix} argmax \\ S_i \in Senses_D\,(w) \end{matrix} P(S_i) \prod_{j=1}^{m} P(f_j|S_i) \qquad (5)$$

In the expressed Eq. 5, features are signified by m and probability *P(Si)* is computed from the frequency metrics in training set. The *P(fj | Si)* is computed from the feature metrics.

CBOW model makes use of neighbouring words to predict probability of current words (Chen et al., 2022). CBOW is termed the reverse version of skip-gram model (Song, 2016). The CBOW likelihood function in Eq. 7 gives probability of most words which appears within a context and the vectors of the context are articulated as the total of the words in the context by the Eq. 6 and Eq. 7(Nakamura & Kimura, 2019):

$$v(c_t) \sum_{u \in c_t} v(u) \qquad (6)$$

$$p(c_t|w_t) = \frac{\exp\,(v(c_t) \times v(w_t)}{\sum_{w\prime \in v} \exp\,((v(c_t) \times v(w'))\prime} \qquad (7)$$

where V is a set of matching words in the dataset. If Eq. 7 is maximized as a likelihood function, then $v(c_t)\ and\ v(w_t)$ are similar.

Sentence-Transformers such as Sentence Bidirectional Encoder Representations Transformers (SBERT) which is a modification of the BERT network that gives an opportunity to compute cosine similarity between sentence embeddings and allow chromosome mapping of sentences into high-dimensional dense vector representations such that sentences with comparable semantic significance are closely matched (Patel et al., 2021). Sentence-transformer framework in python verifies similarity between original sentence and the context (Wahyutama & Hwang, 2022). Word2Vec, FastText, and Global Vectors for Word Representation (GloVe) which are traditional static word embedding approaches do not generate word embedding vectors that indicate the connotation of context (Seo et al., 2022). Sentence-Transformers computes sentence embeddings which are associated through cosine similarity technique to uncover sentences with similar meanings (Ramnarain-Seetohul et al., 2022).

3. Research Methodology

In this research an experiment is conducted using Word2Vec and SBERT to change words into vector shapes in order to discover value of vector nearness between words (Manalu et al., 2019). The experiment is conducted using existing library Gensim as an open-source tool to implement Word2Vec model with python programming language. The research experiment on SBERT, CBOW and Skip-gram architectures to compute vector representations of words. Inputs for similarity measure is a pair of words from the dataset (Manalu et al., 2019).

The corpus is obtained from South African Digital Language Resources (SADiLAR) as unstructured data stored as a text file, which is further filtered and cleaned. The researcher embarked on a process to pre-process the dataset. The researcher build the following pipelines for the pre-processing steps:

Special characters such as punctuation marks (points, commas, question marks, exclamation marks, numeric numbers) and other characters such as ($, %, *, &, etc.) are removed before we start with tokenization (Imaduddin et al., 2019). Each word is tokenized to use it as a proper input, as researcher could not feed a word as a text string into a model. The tokenized dataset is used as a training dataset. Tokenization is a process to split input data strings into wors or tokens (Sumedh Kadam; Aayush Gala; Pritesh Gehlot; Aditya Kurup; Kranti Ghag, 2018).

4. Model Evaluation

The research focus on intrinsic evaluation to assess implementation of word embeddings models. The tactic emphasis on assessing semantic and syntactic relationship among words that are being measured (Phua et al., 2020). Intrinsic evaluation focus on semantic and synctactic relations between words with agregate computation based on correlation coefficient to serve an absolute standard measure (Phua et al., 2020).

The cosine similarity metric adopted in this research determines similarity between data objects from a dataset treated as a vector. Similarity measure is a function to compute the degree of similarity among data objects converted into vectors (Reshma et al., 2020). The mathematical formula below measures the similarity of vectors by the angle amongst two vectors(Tian & Lv, 2018):

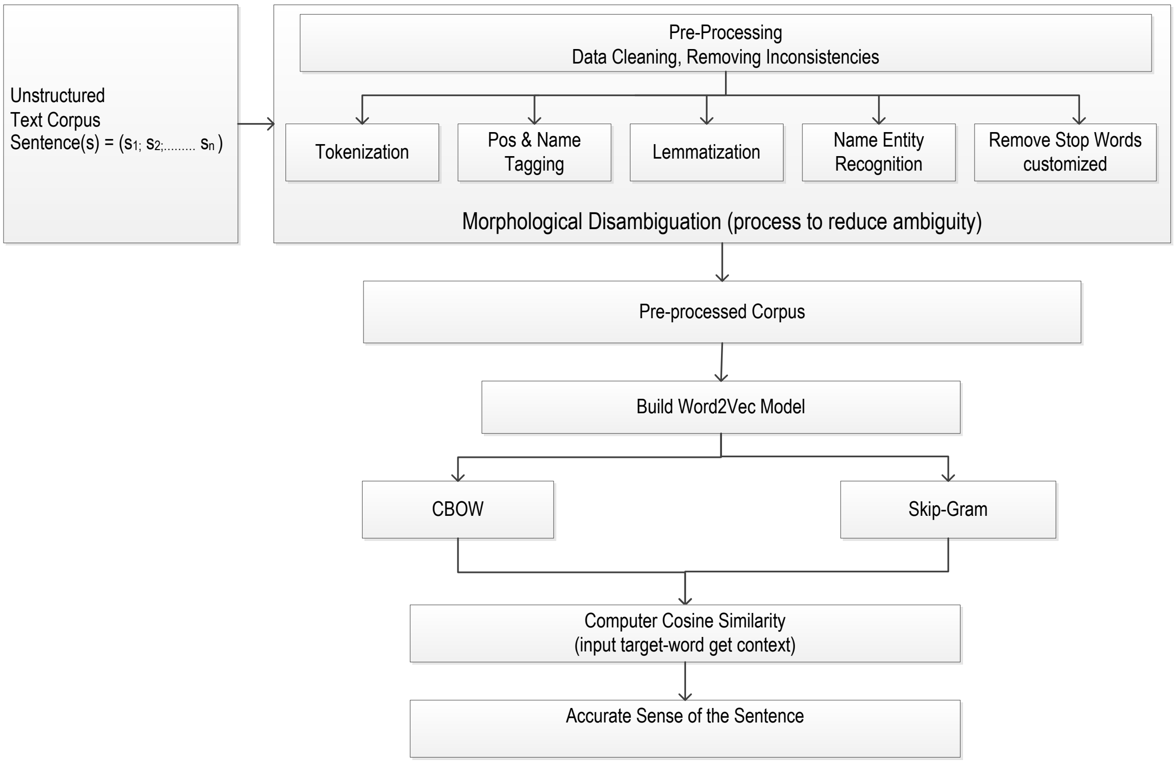
$$\cos(\theta) = \frac{\sum_1^n (A_i \times B_i)}{\sqrt{\sum_1^n A_i^2} \times \sqrt{\sum_1^n B_i^2}} \qquad (8)$$

Sentence Transformers, Eq. 5 for skip-gram and Eq. 7 for CBOW models helps the researcher to express the word in the equivalent vector form, thereafter Eq. 8 is used to compute the cosine similarity to sort output words that are contained by a particular threshold semantically similar to keyword.

5. Research Results

5.1 WSD Pipeline Framework

Figure 2. WSD Pipeline Framework


Unstructured
Text Corpus
Sentence(s) = (s1; s2;........ sn)
Pre-Processing
Data Cleaning, Removing Inconsistencies
Tokenization
Pos & Name Tagging
Lemmatization
Name Entity Recognition
Remove Stop Words customized
Morphological Disambiguation (process to reduce ambiguity)
Pre-processed Corpus
Build Word2Vec Model
CBOW
Skip-Gram
Computer Cosine Similarity
(input target-word get context)
Accurate Sense of the Sentence

The proposed WSD pipeline framework consists of pre-processing pipeline as shown in Fig. 2, for cleaning the text which includes removing punctuation words, numerals, removing customized stop words, and tokenization which splits text into words (Gopal & Haroon, 2016)(Suleiman et al., 2019). The input is a plain unstrucured text, which is pre-processed using the sub-modules for the NLP pipeline. The steps are significant to pre-process the original unstructured dataset as input to the pipeline. The Pre-processing pipeline transform unstructured text into labeled dataset by pre-processing of the input using tokenization, part-of-speech (POS) tagging, stop words removals, lemmatization, multiword detection and named entity recognition recognition (NER) steps (Manalu et al., 2019). The pre-processing help to minimizew ambiguity of words by extracting information from their morphology and context.

The pre-processing pipeline produce a pre-processed corpus, which is then used to build a training Word2Vec model. Continuous bag of words (CBOW) in experiment II (4.3) and skip-gram in experiment I (4.2) are two types of Word2Vec models that train words into word vector space model (Chen et al., 2022). Word2Vec model makes it feasible to compute semantic similarity as it utilizes neural network to learn word associations from the dataset. Word2Vec find the distribution of the expression for a target word by stipulating the context and further makes use of cosine similarity to calculate the similarity actions of the vectors (Reshma et al., 2020). Words with equivalent meanings have a tendency to have the identical word embedding (Manalu et al., 2019).

5.2 Experimental Results CBOW and Skip-Gram Model

Table 1: Words Distribution

# Sesotho sa Leboa Sentences	100964
# Sesotho sa Leboa words	2386351
Old Length	12196163
New Length After Removing stop words	9215488
Vocabulary Size	65272
Vector Size	100
Alpha	0.025

Tab. 2 and Tab. 3 show outputs from an experiment with CBOW and Skip-Gram Word2Vec model which is a popular word-embedding approach that effectively capture semantic and syntactic word similarities from the dataset.

Table 2: Word2Vec – CBOW Model Most Similar word

	Cosine Similarity	**Angle**
scorpion	0.6190912127494812	51.75
kings	0.6004844903945923	53.13
tsopotå¡e	0.5199252367019653	58.73
lehono	0.5128013491630554	59.2
bahlakudi	0.5065954327583313	59.60
dipolitikitå¡a	0.496658593416214	60.26
maphodiksa	0.4800688326358795	61.31

bosenyi	0.46610745787620544	62.54
mpahlwa	0.4615779519081116	62.54
turwa	0.45731163024902344	62.80

Table 3: Word2Vec Skip-gram Model

	Cosine Similarity	**Angle**
thetosello	0.2354934811592102	76.40
nkwane	0.23058441281318665	76.70
nokeng	0.23009748756885529	76.70
sebjaneng	0.22779503464698792	76.87
makgarebe	0.21069371700286865	77.87
masogana	0.20032326877117157	78.46
mahlong	0.1825927495956421	79.51
mabedi	0.17644529044628143	79.86
masea	0.1606215089559555	80.79
maoto	0.16003385186195374	80.79

5.3 Experimental Results Sentence Transformation model using SBERT

Table 4: Corpus with example sentence

Thušo ya bongaka bare chelete yaka efedile gobona ngaka yabasadi
Ngaka ya meno ga epatele meno agodulela ruri kaganong kera goyepa marenini?
Go ntshiwa go ba go phumulwa ga bana mo go thuso ya bongaka ge ba fitile mengwaga ye masome pedi tee.
Ge o oketša palo ya batho go thuso ya bongaka gwa tura, tshelete ya namelela
Tshelete Gago shala mo karateng ya Ngaka, gae fiwe Rena ge e shetje
Mosadi o noka seshebo ka letswai
Noka ya mma e bohloko
Noka e tletse ka meetsi

Table 5: Query Sentences

Makuwa o kgamile ke meetsi, a tshela Noka ya matsekutseku
Thušo ya bongaka e gana ke ntsha meno

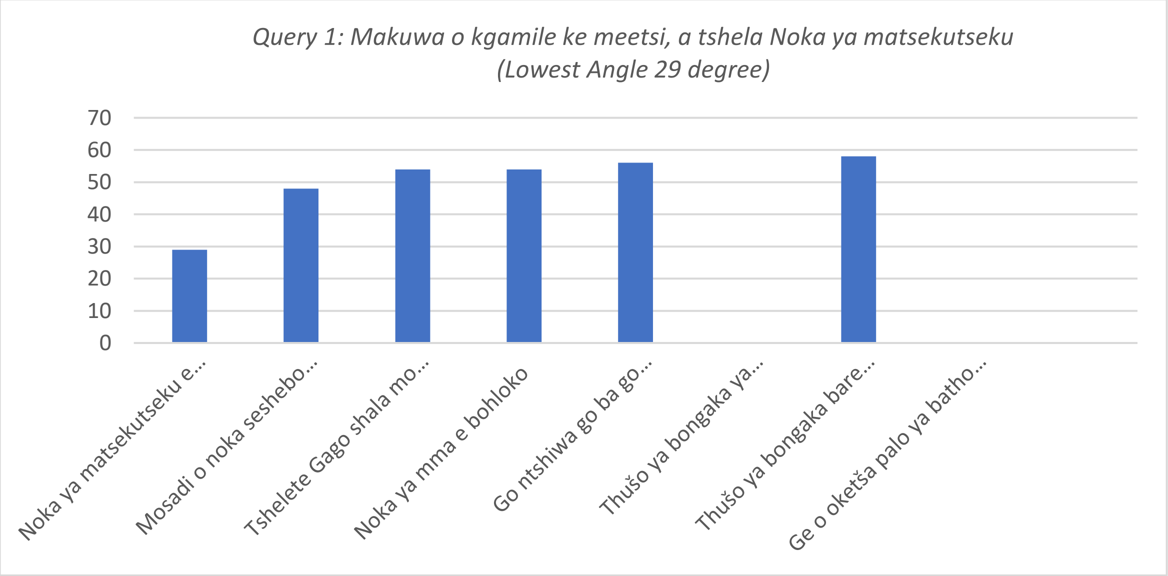
Thušo ya bongaka e ya tura

Table 6: Sentence Transformation with Cosine Similarity

```
tensor([[ 0.0166,  0.0903, -0.0653,  ...,  0.0532,  0.0384, -0.0319],
        [-0.0340,  0.1257, -0.0600,  ..., -0.0005,  0.0213, -0.0111],
        [-0.0244,  0.1753, -0.0683,  ...,  0.0142, -0.0143, -0.0199],
        ...,
        [-0.0426,  0.0721,  0.0075,  ...,  0.0341,  0.0358, -0.0382],
        [-0.0022,  0.0822,  0.0329,  ...,  0.1045, -0.0037, -0.0240],
        [-0.0045,  0.1136,  0.0160,  ..., -0.0249, -0.0235, -0.0544]])
```

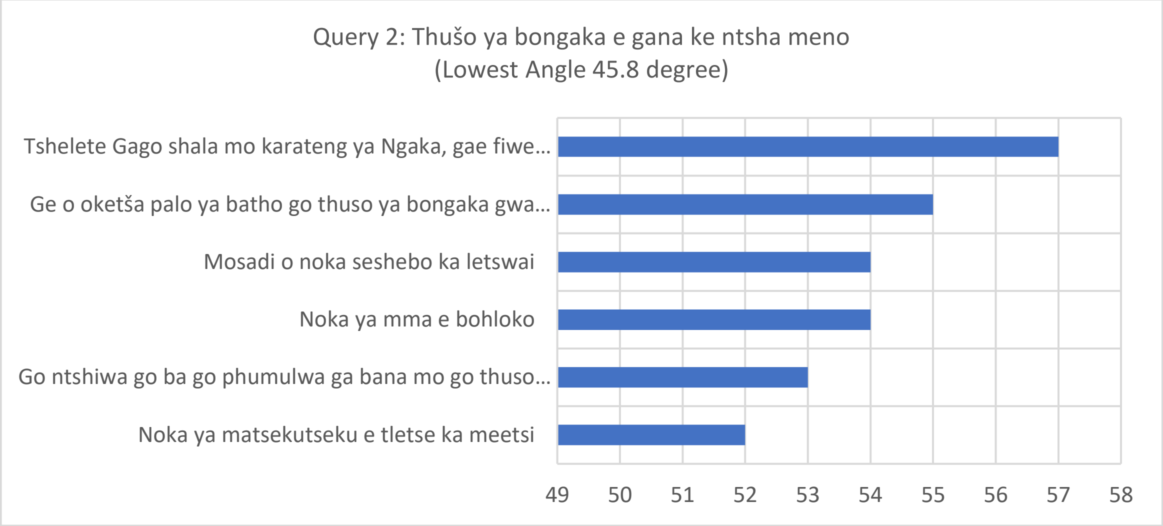
Query 1: Makuwa o kgamile ke meetsi, a tshela Noka ya matsekutseku

Sentence	Cosine Similarity	Angle
Noka ya matsekutseku e tletse ka meetsi	0.872	29
Mosadi o noka seshebo ka letswai	0.662	48
Tshelete Gago shala mo karateng ya Ngaka, gae fiwe Rena ge e shetje	0.584	54
Noka ya mma e bohloko	0.579	54
Go ntshiwa go ba go phumulwa ga bana mo go thuso ya bongaka ge ba fitile mengwaga ye masome pedi tee	0.566	56
Thušo ya bongaka ya meno ga epatele meno agodulela ruri kaganong kera goyepa marenini?	0.56	55.9
Thušo ya bongaka bare chelete yaka efedile gobona ngaka yabasadi	0.526	58
Ge o oketša palo ya batho go thuso ya bongaka gwa tura, tshelete ya namelela	0.503	59.8


Query 1: Makuwa o kgamile ke meetsi, a tshela Noka ya matsekutseku (Lowest Angle 29 degree)

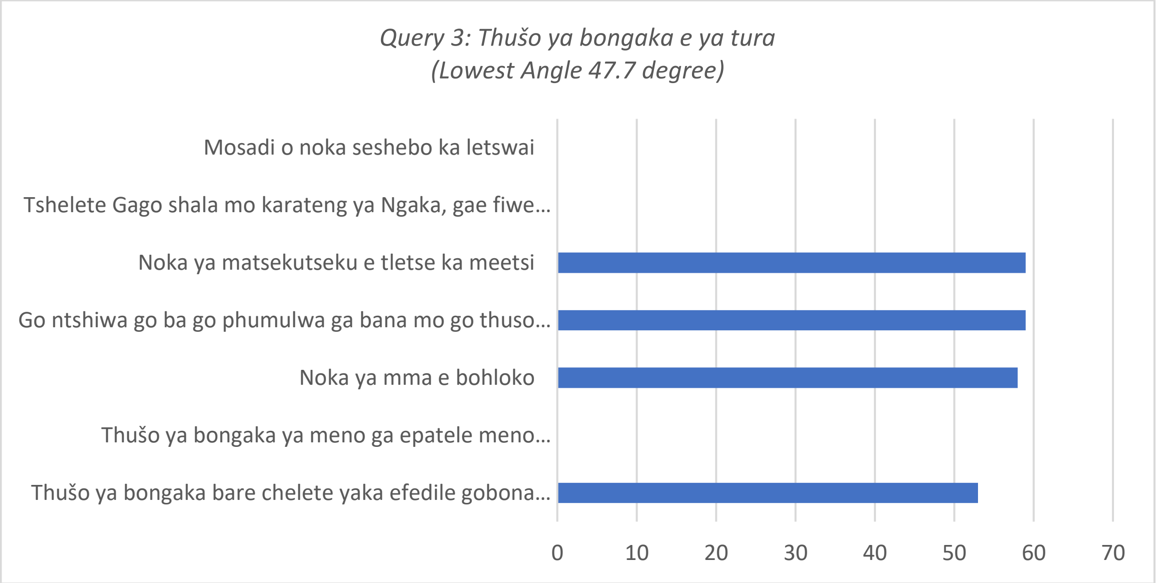
Query 2: Thušo ya bongaka e gana ke ntsha meno

Sentence	Cosine Similarity	Angle
Thušo ya bongaka ya meno ga epatele meno agodulela ruri kaganong kera goyepa marenini?	0.697	45.8
Thušo ya bongaka bare chelete yaka efedile gobona ngaka yabasadi	0.622	51.5
Noka ya matsekutseku e tletse ka meetsi	0.619	52
Go ntshiwa go ba go phumulwa ga bana mo go thuso ya bongaka ge ba fitile mengwaga ye masome pedi tee.	0.608	53
Noka ya mma e bohloko	0.586	54
Mosadi o noka seshebo ka letswai	0.583	54
Ge o oketša palo ya batho go thuso ya bongaka gwa tura, tshelete ya namelela	0.575	55
Tshelete Gago shala mo karateng ya Ngaka, gae fiwe Rena ge e shetje	0.544	57



Query 3: Thušo ya bongaka e ya tura

Sentence	Cosine Similarity	Angle
Ge o oketša palo ya batho go thuso ya bongaka gwa tura, tshelete ya namelela	0.673	47.7
Thušo ya bongaka bare chelete yaka efedile gobona ngaka yabasadi	0.598	53
Thušo ya bongaka ya meno ga epatele meno agodulela ruri kaganong kera goyepa marenini?	0.579	54.6
Noka ya mma e bohloko	0.524	58
Go ntshiwa go ba go phumulwa ga bana mo go thuso ya bongaka ge ba fitile mengwaga ye masome pedi tee.	0.513	59
Noka ya matsekutseku e tletse ka meetsi	0.509	59
Tshelete Gago shala mo karateng ya Ngaka, gae fiwe Rena ge e shetje	0.474	61.7
Mosadi o noka seshebo ka letswai	0.474	61.7



6. Discussion and Conclusion

In this research study, it has been proved that word embedding solve word sense disambiguation (WSD) problems for Sesotho sa Leboa language. To disambiguate words in Sesotho sa Leboa, the Word2Vec model represents the context and each sense of the target word as a vector in a highly dimensional space, and cosine similarity metrics measure semantic relation between the sense definition and context of the ambiguous word. Word2Vec model gives good accuracy with English corpus, as compared to Sesotho sa Leboa corpus, future research will have to optimizes the algorithm and training model for better accuracy on WSD output.

The CBOW model performed well in this research study since the cosine value approach 1, and the angle becomes smaller, which means greater match between vectors giving similar orientation to two vectors. The threshold of CBOW model for this processed text document gives output value for cosine similarity higher than 0.5 which shows strong similarity between context and sense definition.

Skip-gram model gives cosine similarity value closer to 0, with two vectors approaching a perpendicular angle of 90-degrees orthogonally indicating that orientation of vectors do not match. The research results shows no opposite vectors with an angle of 180-degrees, since we do not have a cosine similarity of -1.

Sentence transformer models gives strong similarity setting a new state-of-the art algorithm that outperforms Word2Vec algorithms, as experiment shows a very high cosine similarity of 87%, with an angle of 29 degrees.

In future work, the researchers propose to perform word sense disambiguation using Embeddings from Language Models (ELMo) using Tensorflow-hub for optimization and FastText using Gensim to achieve character and morphological structure of the word.

References

Açıkgöz, O., Gürkan, A. T., Ertopçu, B., Topsakal, O., Özenç, B., Kanburoğlu, A. B., Çam, İ., Avar, B., Ercan, G., & Yıldız, O. T. (2017). All-Words Word Sense Disambiguation for Turkish. *International Conference on Computer Science and Engineering (UBMK)*, 490–495. https://doi.org/10.1109/UBMK.2017.8093442

Agirre, E., & Edmonds, P. (2008). Word Sense Disambiguation. *Scholarpedia*, *3*, 4358. https://www.semanticscholar.org/author/Philip-Edmonds/39934190

Aliwy, A. H., & Taher, H. A. (2019). Word Sense Disambiguation: Survey study. *Journal of Computer Science*, *15*(7), 1004–1011. https://doi.org/10.3844/jcssp.2019.1004.1011

Arksey, H., & O'Malley, L. (2005). Scoping Studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32.

Baˇsi´c, B. D., & ˇSnajder, J. (2018). *Basics of Natural Language Processing Motivation : NLP as preprocessing*. https://www.fer.unizg.hr/_download/repository/TAR-2014-ProjectReports.pdf

Bakx, G. E. (2006). Machine Learning Techniques for Word Sense Disambiguation [Universitat Politμecnica de Catalunya]. In *Machine Learning*. https://www.lsi.upc.edu/~escudero/wsd/06-tesi.pdf

Bangalore, S. (2006). Widening The Nlp Pipeline For Spoken Language Processing. *IEEE Spoken Language Technology Workshop*, 4244. https://doi.org/10.1109/SLT.2006.326787

Bosch, S. E., & Griesel, M. (2017). Strategies for building wordnets for under-resourced languages : The case of African languages. *Literator - Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, *38*(1), 1351. https://doi.org/10.4102/lit.

Bosch, S. E., Jones, J., & PRETORIUS, L. (2007). Towards Machine-Readable Lexicons for South African Bantu languages. *Nordic Journal of African Studies*, *16*(2053403), 131–145.

Daudt, H. M. L., van Mossel, C., & Scott, S. J. (2013). Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. *BMC Medical Research Methodology*, *13*(1), 48. https://doi.org/10.1186/1471-2288-13-48

Dwivedi, R. K., Saxena, A. K., Teerthanker Mahaveer University. College of Computing Sciences & Information Technology, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, & Institute of Electrical and Electronics Engineers. (2019). Vector Representation of Bengali Word Using Various Word Embedding Model. *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 27–30. https://doi.org/10.1109/SMART46866.2019.9117386

FaaB, G. (2010). *A morphosyntacic description of Northern Sotho as a basis for an automated translation from Northern Sotho into English* (Vol. 2010, Issue June) [University of Pretoria]. www.up.ac.za

Flanagan, P. (2013). Adjective Stacking and Classification in Northern Sotho: A Southern Bantu

Language of South Africa. *Lancaster University Postgraduate Conference in Linguistics & Language Teaching 2013*, 1–40.

Giunchiglia, F., Freihat, A. A., & Batsuren, K. (2018). One World – Seven Thousand Languages. *19th International Conference on Computational Linguistics and Intelligent Text*. http://hdl.handle.net/11572/196684

Gopal, S., & Haroon, R. P. (2016). Malayalam word sense disambiguation using Na{\"\i}ve Bayes classifier. *Advances in Human Machine Interaction (HMI), 2016 International Conference On*, 1–4. https://doi.org/10.1109/HMI.2016.7449181

Griesel, M., & Bosch, S. (2012). Taking stock of the African Wordnets project : 5 years of development. In *GWC* (pp. 1–6). https://www.semanticscholar.org/paper/Taking-stock-of-the-African-Wordnet-project%3A-5-of-Griesel-Bosch/e98164265682e3e70f8c713a5dfe44048683f9b0

Imaduddin, H., Widyawan, & Fauziati, S. (2019). Word embedding comparison for Indonesian language sentiment analysis. *Proceeding - 2019 International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019*, 426–430. https://doi.org/10.1109/ICAIIT.2019.8834536

Kulkarni, N., Parachuri, D., Dasa, M., & Kumar, A. (2012). Automated analysis of textual use-cases: Does NLP components and pipelines matter? *Proceedings - Asia-Pacific Software Engineering Conference, APSEC*, *1*, 326–329. https://doi.org/10.1109/APSEC.2012.83

Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: advancing the methodology. *Implementation Science : IS*, *5*(69), 1–9. https://doi.org/10.1186/1748-5908-5-69

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mmaseroka, R. (2006). *POLYSEMY OF THE VERBS YA AND TLA IN BY* (Issue December). University of Stellenbosch.

Navigli, R., & Velardi, P. (2005). Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(7), 1075–1086. https://doi.org/10.1109/TPAMI.2005.149

Pal, A. R., & Saha, D. (2015). WORD SENSE DISAMBIGUATION : A SURVEY. *International Journal of Control Theory and Computer Modeling (IJCTCM)*, *5*(3), 1–16.

Pirkola, A. (2001). Morphological typology of languages for IR. *Journal of Documentation*, *57*(3), 330–348. https://doi.org/10.1108/EUM0000000007085

Popov, A. (2018). Neural Network Models for Word Sense Disambiguation : An Overview. *Cybernetics And Information Technologies*, *18*(1), 139–151. https://doi.org/10.2478/cait-2018-0012

Pretorius, R., & Pretorius, L. (2009). Setswana Tokenisation and Computational Verb Morphology : Facing the Challenge of a Disjunctive Orthography. *EACL Workshop on Language Technologies for African Languages (AfLaT)*, *March*, 66–73.

Puttkammer, M. J., Eiselen, E. R., Hocking, J., & Koen, F. J. (2018). NLP Web Services for Resource-Scarce Languages. *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics-System Demonstrations, 43–49.

Rahab, N., & Mothapo, B. (2019). *Determining the core vocabulary used by Sepedi- speaking preschool children during regular preschool-based activities by* (Issue April) [University of Pretoria]. www.up.ac.za

Ranjan, J., & Ranjan, R. (2010). *Papers Selected From International Conference On Innovation In Redefining Business Horizons Institute of Management Technology , Ghaziabad , India , 18 - 19 December , Application of Data Mining Techniques In Higher Education In India*. *11*(1), 18–19.

Shallu, & Gupta, V. (2013). A survey of word-sense disambiguation effective techniques and methods for Indian languages. *Journal of Emerging Technologies in Web Intelligence*, *5*(4), 354–360. https://doi.org/10.4304/jetwi.5.4.354-360

Sumedh Kadam; Aayush Gala; Pritesh Gehlot; Aditya Kurup; Kranti Ghag. (2018). Bayes Algorithm for Spam Filtering. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1–5. https://doi.org/10.1109/ICCUBEA.2018.8697601

Sutor, J. P., Fermuller, C., Summers-stay, D., & Aloimonos, Y. (2019). Metaconcepts : Isolating Context in Word Embeddings. *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 1–6. https://doi.org/10.1109/MIPR.2019.00110

Tasnim, Z., Ahmed, S., Rahman, A., Sorna, J. F., & Rahman, M. (2021). Political ideology prediction from bengali text using word embedding models. *2021 International Conference on Emerging Smart Computing and Informatics, ESCI 2021*, 724–727. https://doi.org/10.1109/ESCI50559.2021.9396875

Tomar, A., Bodhankar, J., Kurariya, P., Anarase, P., Jain, P., Lele, A., Darbari, H., & Bhavsar, V. C. (2013). High performance natural language processing services on the GARUDA grid. *2013 National Conference on Parallel Computing Technologies (PARCOMPTECH)*, *February*, 1–6. https://doi.org/10.1109/ParCompTech.2013.6621407

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, J. D. (2013). Distributed Representations ofWords and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Neurips. https://arxiv.org/abs/1310.4546?context=stat

Torii, M., Fan, J. W., & Zisook, D. S. (2015). Finding Difficult-to-Disambiguate Words: Towards an Efficient Workflow to Implement Word Sense Disambiguation. *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*, 448. https://doi.org/10.1109/ICHI.2015.66

Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C.-C. J. (2019). Evaluating Word Embedding Models: Methods and Experimental Results. *APSIPA Transactions on Signal and Information Processing*, *8*(e19), 13. https://doi.org/10.1017/ATSIP.2019.12