SKIP-GRAM BASED GRAMMAR CORRECTOR USING SEMANTIC AND SYNTACTIC ANALYZER FOR NEPALI

Archit Yajnik

Department of Mathematics, Sikkim Manipal Institute of Technology, Sikkim, India

ABSTRACT

Unlike English, Grammar checker is a vital problem for many languages in India. The Grammar corrector (GC) based on the syntactic and semantic information of a Nepali sentence is modelled. Skip-gram model is used for the word to vector encoding. Window size of 3 context words is employed for the word to vector encoding. The network is trained up to the negative log entropy goes to 0.05. The network is tested over 500 incorrect syntactics and semantics of Nepali sentences. The network has suggested the corrections with the accuracy of 96.4%.

KEYWORDS

Skip-Gram, Grammar Corrector, word embedding

1. INTRODUCTION

The Empirical evaluations of existing grammar checkers have also gained prominence in recent years. In 2020, Subham Sahu and his team assessed five popular grammar-checking applications—Grammarly, Ginger, ProWritingAid, LanguageTool, and After the Deadline—using a dataset of 500 sentences. These sentences were evenly distributed among five main grammar error categories: sentence structure, punctuation, spelling, syntax, and semantics. The study revealed that Grammarly achieved the highest overall accuracy at 44.4%, while After the Deadline recorded the lowest accuracy at 28.74%. Notably, none of the applications scored above 11% for sentence structure errors, highlighting significant room for improvement in handling complex grammatical issues. Rules on Grammar checker is presented by Miłkowski, M in 2012.

Further advancements in 2023 included the development of a grammar checker for the Yorùbá language, which utilized Government and Binding Theory to address unique syntactic structures. Although this research contributed to linguistic inclusivity, it did not specify dataset sizes or accuracy rates. Additionally, Ronald Schmidt-Fajlik (2023) examined ChatGPT as a grammar checker for Japanese English learners, comparing its performance with traditional tools like Grammarly and ProWritingAid. His study suggested that AI-driven models provided contextually relevant feedback but still faced challenges in comprehending nuanced linguistic errors. Similarly, Robert Long (2022) investigated the efficacy of online grammar checkers versus self-editing among English as a Foreign Language (EFL) students. His findings indicated only marginal differences between the two approaches, reinforcing the notion that while grammar checkers are useful, they should complement rather than replace human editing.

David C. Wyld et al. (Eds): NLMLT, SEAU, ITCS, CST – 2025
pp. 15-21, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.151602

User perception studies have also played a crucial role in evaluating grammar checkers. Fatma Yurika et al. (2023) explored student perceptions of Grammarly Premium during academic writing. Their research revealed that students appreciated the accessibility and detailed feedback provided by Grammarly but raised concerns about occasional inaccuracies and over-reliance on the tool. Similarly, Dila Anggita and colleagues (2023) conducted a systematic review to assess the effectiveness of online grammar checkers in supporting EFL learners. Their findings suggested that while these tools offer significant benefits, they cannot fully replace the nuanced feedback provided by human instructors. The literatures [6], [8] and [9] demonstrate the computational advancement in grammar of Nepali.

2. CORPUS DETAILS

In Indo-Aryan languages matras or case marker (Vibhakti) are extremely important in a sentence, any mistake in it will change the meaning of the sentence and many times sentence becomes illogical and incorrect. Few such sentences are highlighted in table 2. These are the common mistakes a person makes while framing the sentences [7].

Table 1. Spelling Error

Sr. No	Nepali Words with Spelling Error	Reference Spelling	Gloss
1	मेरो घर **खालि** छ खालि - Only खाली - Empty	मेरो घर **खाली** छ	My house is empty
2	तपाईंमा मेरो हार्दिक **संवेदना** छ संवेदना – Emotions समवेदना - Condolence	तपाईंमा मेरो हार्दिक **समवेदना** छ	My heartfelt condolence is with you

The article emphasis on checks the spelling (हिज्जे in Nepali and वर्तनी in Hindi) for the given Nepali sentence. Semantic and syntactically correct 680 sentences are collected in the corpus.

3. METHODOLOGY

Skip-Gram is briefly described as depicted in figure #

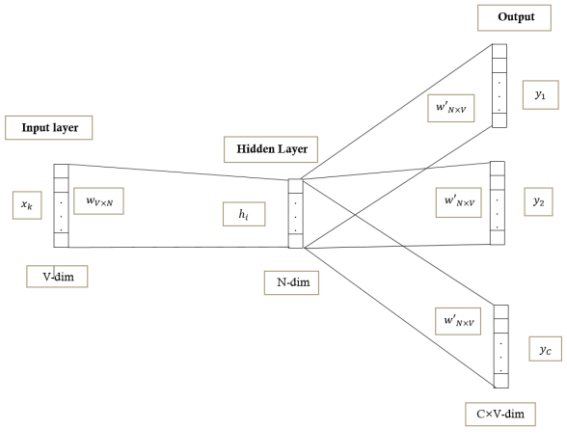

Figure 1: Skip-Gram Neural Network model in general

In the Figure 1 above,

$w_i = word\ i\ from\ vocabulary\ V$

v$\in \mathbb{R}^{n\times|V|}$ $: Input\ word\ matrix$

$v_i : i-th\ column\ of$ v, the input vector representation of word w_i

u $\in \mathbb{R}^{|V|\times n}$: Output word matrix

$u_i : i-th\ row\ of$ v, the input vector w_i

Forward Part

We generate our one-hot input vector of the centre word

X $\epsilon\ \mathbb{R}^{|V|}$

We get our embedded word vector for the centre word using the matrix V

VC= v X $\epsilon\ \mathbb{R}^{n}$

We generate a score vector by multiplying the matrix U by the embedded vector

Z= μVC

Turn the score vector into probabilities using softmax function

$\hat{y} = softmax(z)\ \epsilon\ \mathbb{R}^{|V|}$

$\hat{y}_{C-m,\ldots,}\hat{y}_{C-1,}\hat{y}_{C+1,\ldots,}\hat{y}_{C+m,}$

This is how input are propagated forward through the network to produce output.

The probability vector generated should match the true probabilities (context word) of the one-hot vectors of the actual output.

$y^{(c-m)},\ldots,y^{(c-1)},y^{(c+1)},\ldots,y^{(c+m)}$

$\hat{y} = softmax(z_j)\ \epsilon\ \mathbb{R}^{|V|} = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}}$, for j = 1,2,…,k

The probabilities of observing each context word is,

$\hat{y}_{C-m,}\ \ldots\ ,\ \hat{y}_{C-1},\hat{y}_{C-1,C+1}\ ,\ \ldots\ ,\ \hat{y}_{C+m,}$

Minimize E = -log P $(w_{c-m},\ldots,w_{c-1},w_{c+1}\ \ldots,w_{c-m}|w_c)$

= - log $\prod_{j=0,j\neq m}^{2m}$ P $(w_{c-m+j}\ |w_c)$ (using unigram model)

$$=- \log \prod_{j=0, j\neq m}^{2m} P\left(u_{c-m+j} \mid v_c\right) \qquad \text{(writing it in vector form)}$$

Using the Softmax function, which is used to convert the raw scores into probabilities

$$P\left(u_{c-m+j} \mid v_c\right) = \frac{exp\left(u_{C-m+j}^T \mid v_c\right)}{\Sigma_{k=1}^{|V|} exp\left(u_k^T v_c\right)}$$

Therefore, the loss function becomes

$$E = - \log \prod_{j=0, j\neq m}^{2m} \frac{exp\left(u_{C-m+j}^T \mid v_c\right)}{\Sigma_{k=1}^{|V|} exp\left(u_k^T v_c\right)}$$

By simplifying it, we get

$$E = -\Sigma_{j=0, j\neq m}^{2m} \frac{exp\left(u_{C-m+j}^T \mid v_c\right)}{\Sigma_{k=1}^{|V|} exp\left(u_k^T v_c\right)}$$

$$E = -\Sigma_{j=0, j\neq m}^{2m} u_{C-m+j}^T \; v_c + 2m \log \Sigma_{k=1}^{|V|} exp(u_k^T v_c)$$

Which is simply the probability of the output words (the words in the input word's context) given the input word.

Now, let's derive the gradient with respect to the target word embedding v_c and the context word embeddings u.

Gradient with respect to v_c

The loss function in terms of v_c is:

$$E = - \Sigma_{j=0, j\neq m}^{2m} u_{C-m+j}^T \; v_c + 2m \log \Sigma_{k=1}^{|V|} exp(u_k^T v_c)$$

$$\text{Or, } E = -\Sigma_{j=0, j\neq m}^{2m} \left(u_{C-m+j}^T \; v_c - \log \Sigma_{k=1}^{|V|} exp(u_k^T v_c)\right)$$

Taking the partial derivative with respect to v_c:

$$\frac{\partial E}{\partial v_C} = -\Sigma_{j=0, j\neq m}^{2m} \left(u_{c-m+j} v_c - \frac{\Sigma_{k=1}^{|V|} u_k \, exp\left(u_k^T v_c\right))}{\Sigma_{k=1}^{|V|} exp\left(u_k^T v_c\right)}\right)$$

$$\frac{\partial E}{\partial v_C} = - \Sigma_{j=0, j\neq m}^{2m} \left(u_{c-m+j} v_c - \Sigma_{k=1}^{|V|} P(u_k|v_C)\, u_k\right)$$

Where, $P(u_k|v_C)\, u_k) = \frac{exp\left(u_k^T v_c\right))}{\Sigma_{k=1}^{|V|} exp\left(u_k^T v_c\right)}$

$$\frac{\partial E}{\partial v_C} = - \Sigma_{j=0, j\neq m}^{2m} u_{c-m+j} v_c + 2m \, \Sigma_{k=1}^{|V|} P(u_k|v_C)\, u_k$$

From backpropagation we know that,

$$w'(new) = w'(old) - \eta \frac{dE}{dw'}$$

Finally we arrive at our gradient descent equation for our input weights

$$v_C \longleftarrow v_C - \eta(- \Sigma_{j=0, j\neq m}^{2m} u_{c-m+j} v_c + 2m \Sigma_{k=1}^{|V|} P(u_k|v_C)\, u_k)$$

Where, η = learning rate

4. RESULT AND DISCUSSION

The loss graph occurred during training is depicted in in fig (loss). the network is trained using 680 sentences correct sentences and converged with the error goal 0.05.

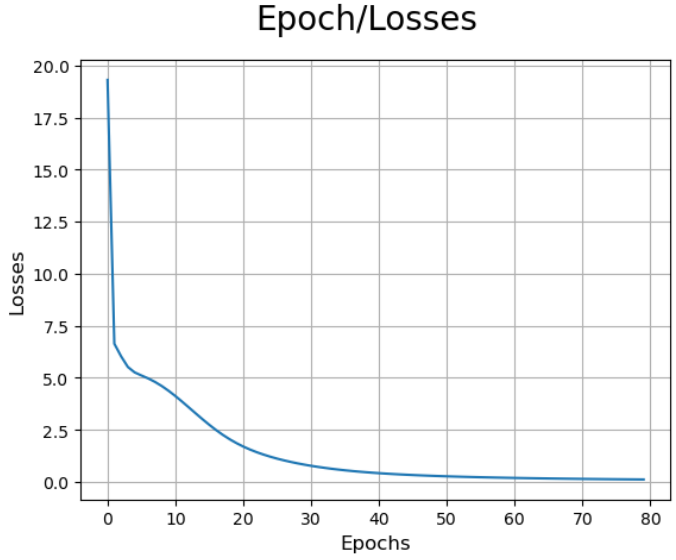

Figure 2. Loss graph of Skip Gram

The outcome of GC is illustrated with following examples while given wrong sentences as an input.

Sentence : I दाजु स्कुल जान्छिन्।

Error: दाजु is masculine and जान्छिन् is Feminine
Correction suggested:

जान्छन
Correct sentence is दाजु स्कूल जान्छन
दिदी
Correct sentence is दिदी स्कूल जान्छिन

Sentence: 2 दिदी रोटी खाइनन् |

Error: दिदी does not have appropriate post position (Vibhakti)

Correction suggested:

दिदीले
Correct sentence is दिदीले रोटी खाइनन्

Sentence: 3 बहिनी पुस्तक पढिन्छ |

Error: Due to the presence of पढिन्छ , the Sentence is in passive voice but the sentence is written in active voice.

Correction suggested:

पढ्छे

Correct sentence is बहिनी पुस्तक पढ्छे
बहिनीद्वारा

Correct sentence is बहिनीद्वारा पुस्तक पढिन्छ

5. CONCLUSIONS

Furthermore the wrong sentences are directed to GC for necessary corrections and suggestions. The size of the context window is considered in this article is 3 for Skip Gram. The experiment is focused on the most prevalent orthography error while writing Nepali , i.e.

1. inappropriate verb with respect to the gender of the subject (Karta)
2. missing case marker
3. plural noun is assigned singular verb.

Testing is carried out on 500 projective incorrect syntactic and semantic Nepali sentences, out of which 482 are correctly suggested by GC and achieved 96.4% accuracy.

REFERENCES

[1] Miłkowski, M, (2012) "Automating rule generation for grammar checkers", *Explorations Across Languages and Corpora.* pp 123-133.

[2] Soni, M., & Thakur, J. S. (2018) "A Systematic Review of Automated Grammar Checking in English Language" *arXiv preprint arXiv:1804.00540*.

[3] Wang, Y., Wang, Y., Liu, J., & Liu, Z. (2020) "A Comprehensive Survey of Grammar Error Correction" *arXiv preprint arXiv:2005.06600.*

[4] Naghshnejad, M., Joshi, T., & Nair, V. N. (2020) "Recent Trends in the Use of Deep Learning Models for Grammar Error Handling" *arXiv preprint arXiv:2009.02358.*

[5] Sahu, S., Vishwakarma, Y. K., Kori, J., & Thakur, J. S. (2020) "Evaluating Performance of Different Grammar Checking Tools" *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9, No. 2, pp 2227-2233.

[6] Pradhan, A., & Yajnik, A. (2024) "Parts-of-Speech Tagging of Nepali Texts with Bidirectional LSTM, Conditional Random Fields, and HMM" *Multimedia Tools and Applications*, Vol. 83, pp 9893–9909.

[7] Krishna maya manger, (2018) "Study on Lexico-Semantic Ambiguity, Journal Of Advanced Linguistic Studies", Vol. 7, No. 1-2.

[8] Archit Yajnik and Sabu Tamang, (2023) "Parser based sentiment analysis for Nepali text", *NLPTT*, pp. 1-12

[9] Archit Yajnik and Sabu Tamang, (2023) "Chunker based sentiment analysis and tense classification for Nepali text" *International Journal of Natural Language Computing*, Vol. 12, No. 6.