Nepali Context-Aware Spelling Tool

Team Members:

Anish Raj Manandhar [THA077BCT010]

Nabin Shrestha [THA077BCT026]

Prayush Bhattarai [THA077BCT035]

Supervised By:
Er. Shanta Maharjan

Co-Supervised by:
Er. Prabin Acharya

Department of Electronics and Computer Engineering
Thapathali Campus

August, 2024

Presentation Outline

- Motivation
- Objectives
- Scopes
- Applications
- Methodology
- Results
- Analysis/Discussion of Results
- List of Remaining Tasks
- References

Motivation

- Gap in Research and Development
- Limited Resources
- Inadequate Existing Tools for Contextual Solutions

Objectives

- To develop spell checker that uses sentence context to detect and correct spelling errors.

Scopes

- Data Collection and Preprocessing for downstream tasks
- Contextual Error Detection

Applications

- Media and Publishing
- OCR Projects
- Reliable TTS Systems
- Search Engines
- Text Processor Systems

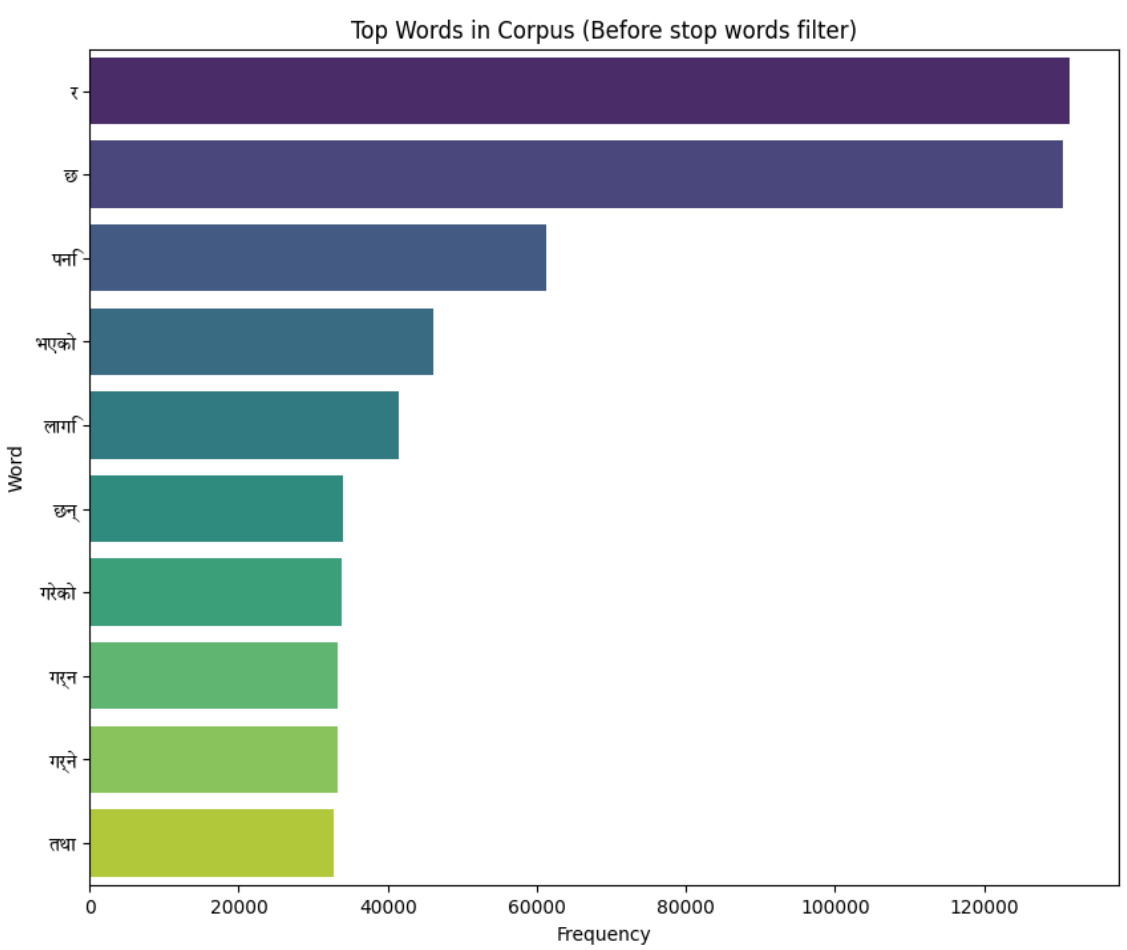
Dataset-[1]: Description

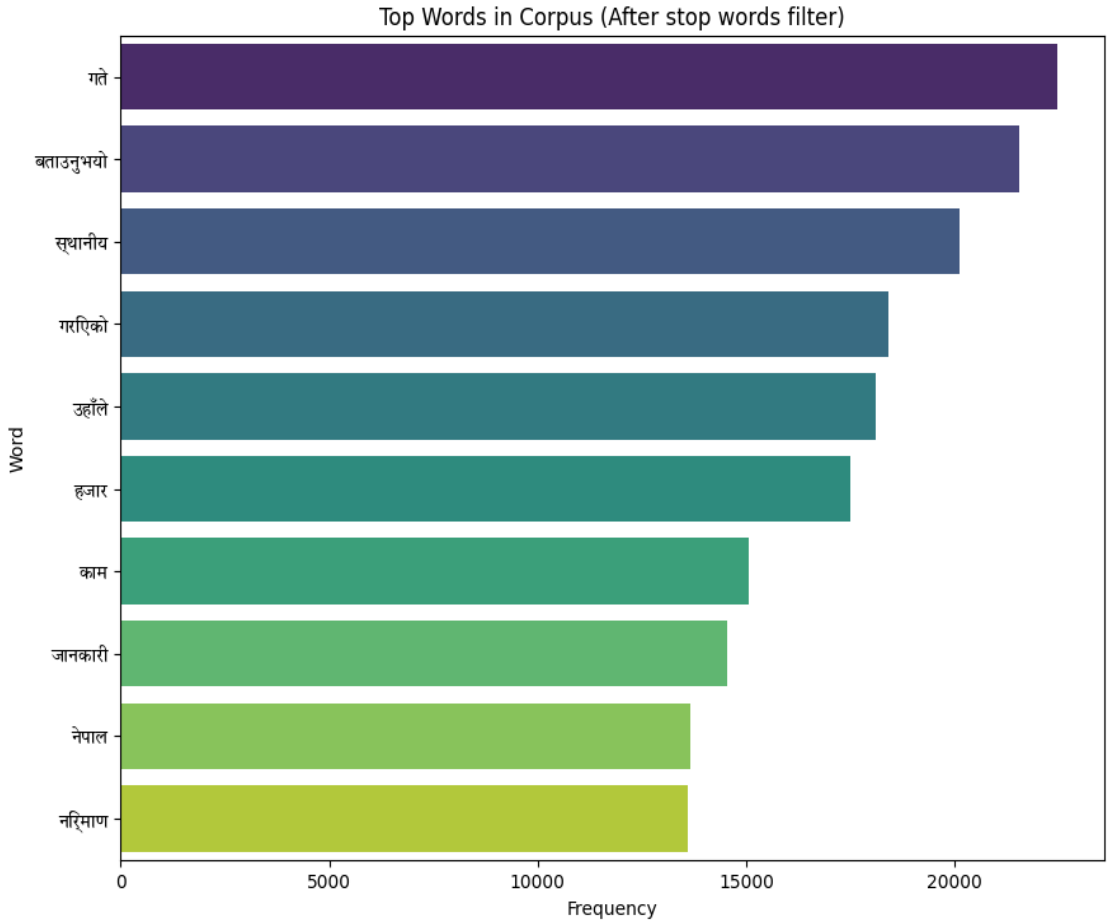
- Collection of Nepali news articles categorized into 20 distinct categories
- extracted from the most trusted Nepali newspapers, such as Kantipur and Gorkha Patra
- 73,000 newspaper articles

	Preprocessing	
Corpus	Before	After
Number of Words	69,61,006	44,70,401
Number of Vocabulary	2,07,458	
Number of sentences	4,92,872	

Category	Number of docs
Agriculture	200
Automobiles	246
Bank	617
Blog	259
Business	307
Economy	600
Education	185
Employment	304
Entertainment	634
Health	180
Interview	330
Literature	251
Migration	111
Opinion	500
Politics	550
Society	353
Sports	700
Technology	118
Tourism	265
World	313

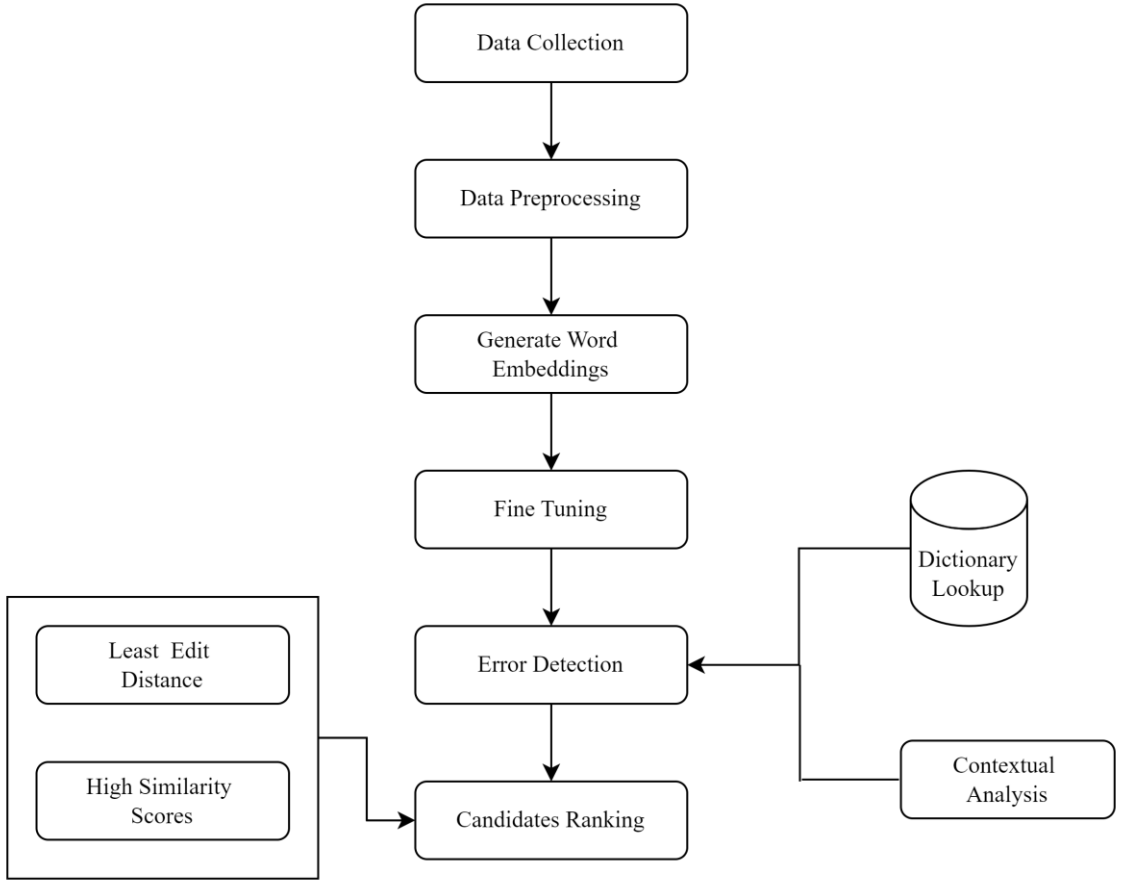
Dataset[2]-Result of preprocessing





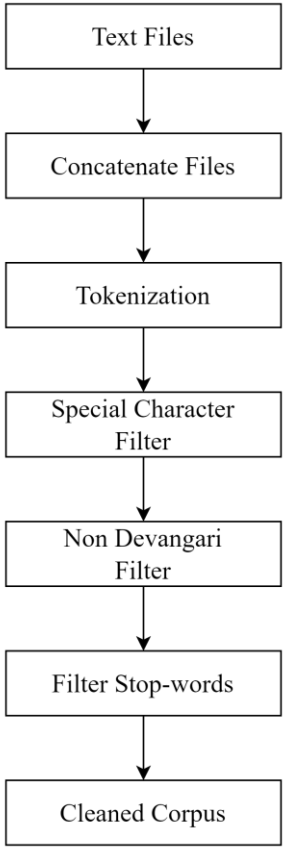
Dataset[3]: Word Cloud

Methodology-[1]
System Block Diagram



Methodology-[2] (Preprocessing pipeline)

- Tokenizer
  - Sentence tokenize
  - Word tokenize
- Special Character Filter
  - Filter characters not used in Nepali
  - Eg:←□…¬ = > < @ # $ ^ & * | \ / ` ~ _ { } [ ]
- Non-Devanagari Filter
  - Regular Expression(Unicode)
- Filter Stop Words
  - Number of stop words filtered: 255



Methodology-[3] (Examples)

- Input: "हार धुनुहोस्, स्वस्थ रहनुहोस्"
- Meaning: "Wash the necklace, stay healthy"
- Issue: Contextually incorrect

Methodology-[4]
(Word Splitting)

- Split sentence into words:
- ['हार’,'धुनुहोस्’, 'स्वस्थ’, 'रहनुहोस्’]

- Store Context Words:
  - 'हार’ → ['धुनुहोस्’, 'स्वस्थ’, 'रहनुहोस्’]
  - 'स्वस्थ’ → ['हार’,'धुनुहोस्’, 'रहनुहोस्’]

Methodology-[5]
(Contextual Similarities Score)

- Calculated Contextual Similarities:
    - 'हार': 0.2
    - 'धुनुहोस्': 0.4
    - 'स्वस्थ': 0.67
    - 'रहनुहोस्': 0.56
- Lowest score: 'हार' (0.2)

Methodology-[6] (Candidate Generation)

- For 'हार’:
- Candidates:
  - [('हार', 0), ('हजार', 1), ('हात', 1), ('हाल', 1), ('हाफ', 1), ('हतार', 1)]
- Filtered Candidates:
  - [(‘हात’, 1, 0.59), ('हारे', 1, 0.614), ('हारि', 1, 0.393), (‘हार', 0, 0.2)]

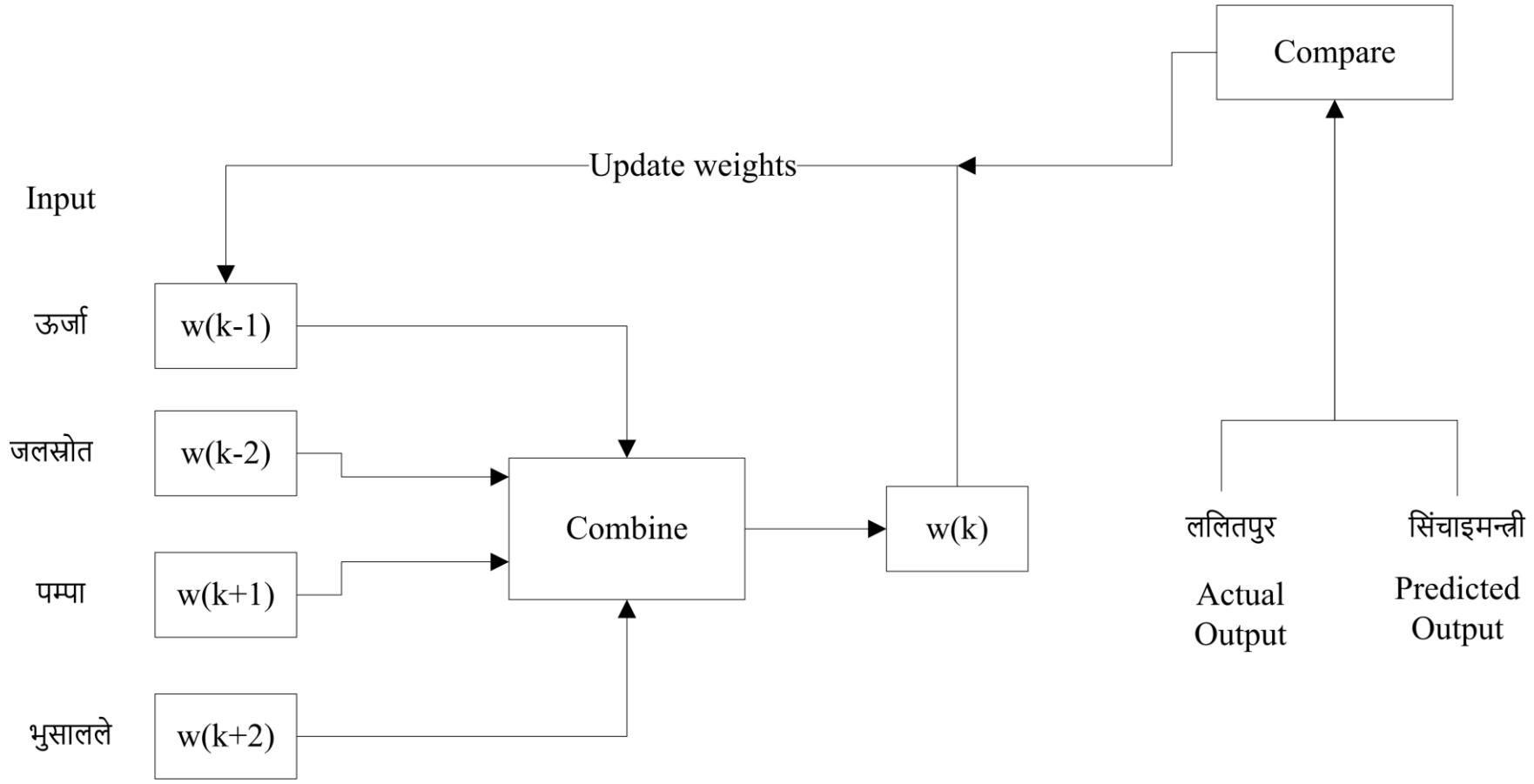
Methodology-[7] (Correction)

- Select Best Candidate
  - Best match: 'हात'
  - Replace 'हार' with 'हात’
- Corrected Sentence
  - Result: "हात धुनुहोस्, स्वस्थ रहनुहोस्"
  - Meaning: "Wash your hands, stay healthy"
  - Correction: 'हार' (necklace) → 'हात' (hand)

Methodology-[8]
(Word2Vec Model)

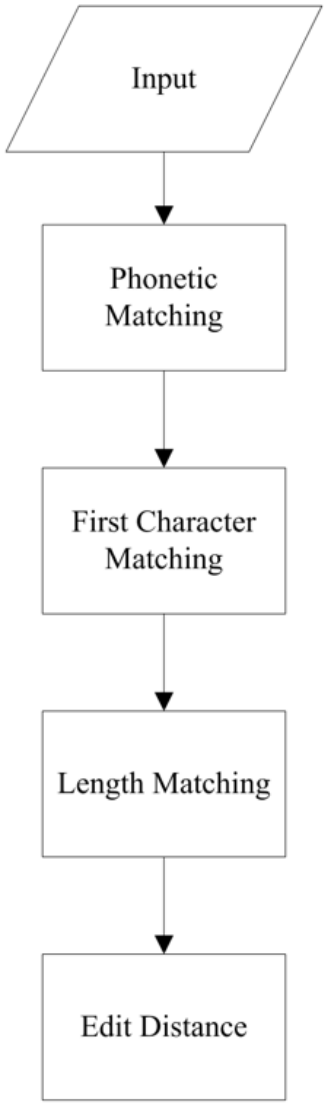
- Transforms words into high-dimensional vector representations.
- Captures semantic relationships between words.
- Similar meanings are located close to each other in vector space.
- Vector representations capture meanings based on context.
- Training on collected corpus.
- Examples: "सुन्दर" and "राम्री"

Methodology-[9] (Word2Vec Architecture)



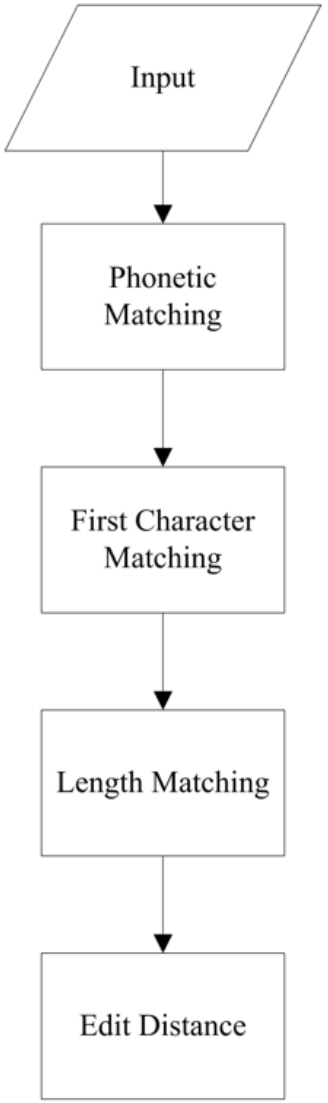
Methodology-[10] (Generate Candidates)

- Phonetic Matching:
  - Convert the erroneous word and each word in the vocabulary into their phonetic codes.
- First Character Matching:
  - Ensure that the first character of the erroneous word and candidate words match.



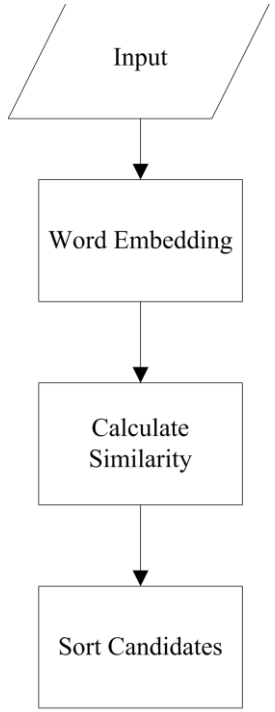
Methodology-[10] (Contd) (Generate Candidates)

- Length Matching:
  - Ensure the lengths of the erroneous word and candidate words match.
- Edit Distance:
  - Calculate the edit distance between the erroneous word and each candidate word.
  - Then, add the candidate word to the list of possible corrections.



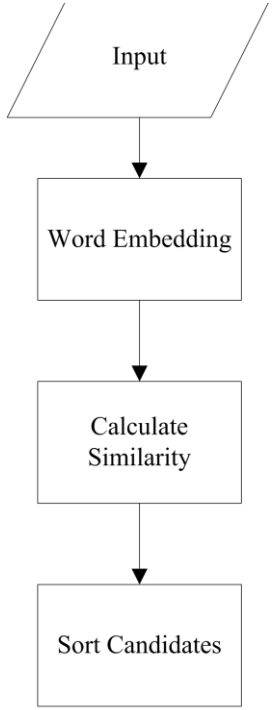
Methodology-[11] (Contextual Filtering)

- Word Embeddings:
  - Use vector representations for each candidate word and context word.



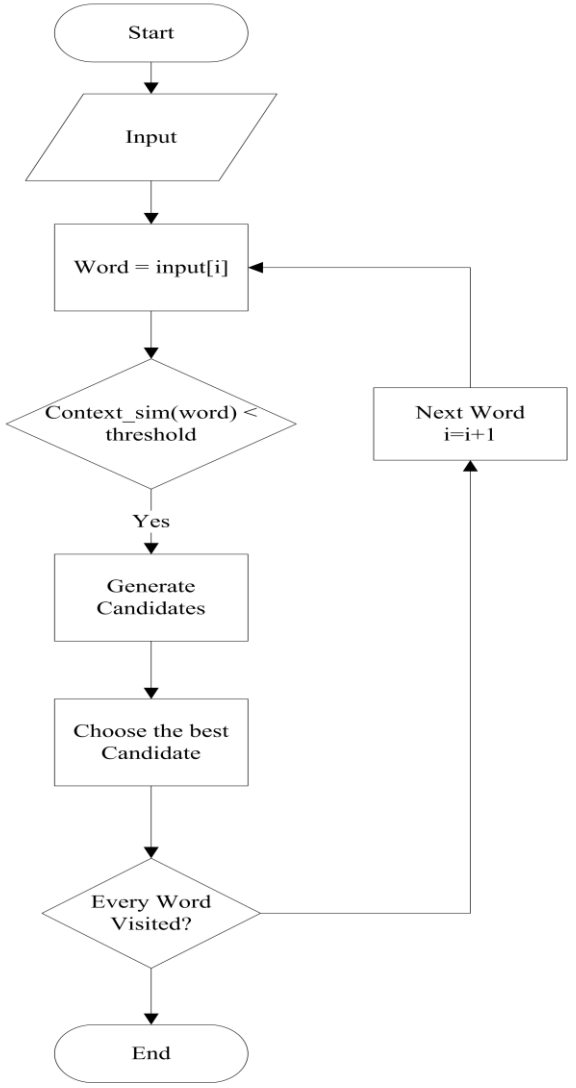
Methodology-[11] (Contd) (Contextual Filtering)

- Calculate Similarity:
  - Calculate the average similarity between its embedding and the embeddings of the context words.
- Sort Candidates:
  - Sort the candidate words based on their similarity scores and edit distances.
  - Highest similarity score and the lowest edit distance are selected.

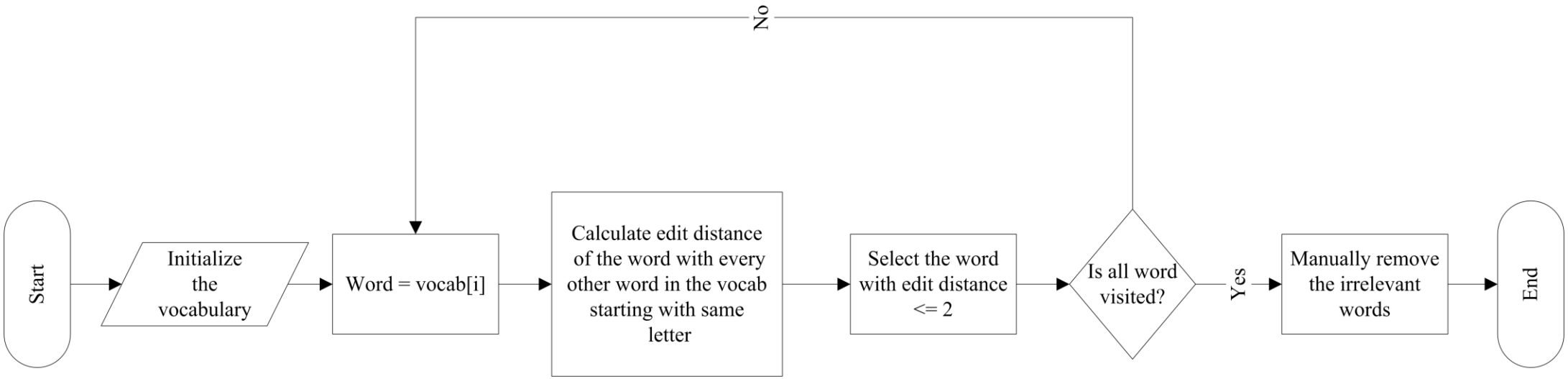


Methodology-[12] (Correct Sentence)

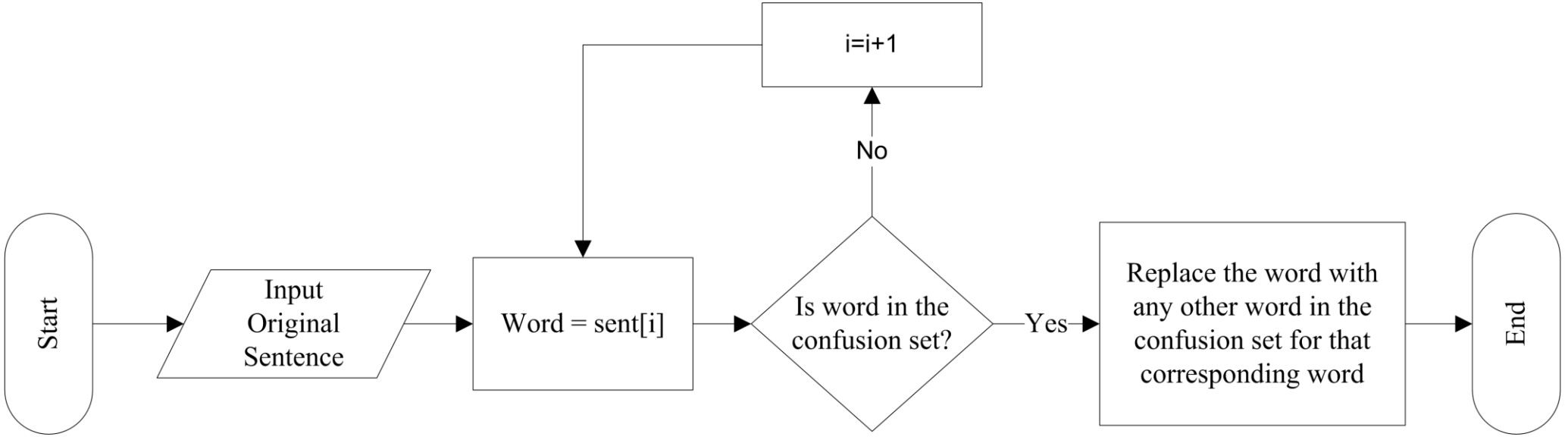
- For each word in the sentence:
  - If the word has a low context similarity score, generate candidate corrections.
  - Filter the candidates based on their context similarity scores.
  - Choose the best candidate as the corrected word.
  - If the word has a high context similarity score, keep it as is.



Methodology-[13]
(Confusion Sets Generation)



Methodology-[14]
(Test Set Generation)



Parameters Specification

Dimension	300
Architecture	CBOW
Epochs	15
Window	10
Minimum count	2
Negative Sampling	15

Results-[1]

Words similar to 'नुन':
दाल: 0.9016
नून: 0.8774
दाना: 0.8659
अन्न: 0.8657
बिस्कुट: 0.8591
तोरीको: 0.8590
रक्सी: 0.8577
चाउचाउ: 0.8543
मिसाएर: 0.8532
चिउरा: 0.8524

Words similar to 'आइतबार':
बुधबार: 0.9195
सोमबार: 0.9191
बिहीबार: 0.9125
शुक्रबार: 0.9104
मङ्गलबार: 0.9033
शनिबार: 0.8834
मंगलबार: 0.8438
सोमवार: 0.8023
शुक्रवार: 0.7719
शनिवार: 0.7700

Words similar to 'जंगल':
जङ्गल: 0.8666
घना: 0.8650
आसपासमा: 0.8510
आसपासको: 0.8407
आसपास: 0.8312
वरिपरिको: 0.8257
तालतलैया: 0.8241
जङ्गलले: 0.8233
किनार: 0.8225
जङ्गलको: 0.8217

Results-[2]

```
Input sentence: उनीहरु  फुल टिप्छु
Candidates : [('उनीहरु', 0), ('उनीहरू', 1), ('उनीहर', 1), ('उनिहरु', 1), ('उनीरु', 1), ('उनीहरुल', 1), ('उनीअरु', 1), ('उनहरु', 1), ('उनीहरुक', 1)]
Filtered Candidates: [('उनीहरुल', 1, 0.10958116129040718), ('उनीअरु', 1, 0.07448022440075874), ('उनीरु', 1, 0.03889932855963707), ('उनिहरु', 1, 0.
Candidates : [('फुल', 0), ('फूल', 1), ('फल', 1), ('फेल', 1), ('फुड', 1), ('फुट', 1), ('फुले', 1), ('फिल', 1), ('फुस', 1), ('फजुल', 1), ('फसल',
Filtered Candidates: [('फूल', 1, 0.2546461895108223), ('फुल्', 1, 0.23844227730296552), ('फुल', 0, 0.22110747545957565), ('फुलि', 1, 0.211271580
Candidates : [('टिप्छु', 0), ('टिप्रु', 1), ('टिप्छ', 1), ('टिप्छन', 1), ('टिक्छु', 1), ('टिप्छे', 1)]
Filtered Candidates: [('टिप्छन', 1, 0.1981523036956787), ('टिप्छु', 0, 0.1326882727444172), ('टिप्रु', 1, 0.11359601840376854), ('टिप्छ', 1, 0.013859424
```

Discussion-[1]

- Relatedness Set
  - Kitchen: रोटी, तरकारी, भित्र, नुन, मसला, अदुवा, लसुन, तेल, मर्चा, दाल, चामल, कराइ, भाडो, पिठो, डाडु, पन्यू, कचौरा, ग्लास
  - Nature: हिमाल, पहाड, जंगल, गन्तव्य, झरना, बाटो, जंगल, चोटी, यात्रा, हिउँ, हरियाली, देउराली, ताल

Discussion-[2]

- Sentiment Set:
  - Positive: राम्रो, सस्तो, जाँगरिलो, नामी, सपना, हर्षोल्लास, सफल, राम्रो, चाँडो, खुशी, बलियो, असल, सुन्दर, सक्षम
  - Negative: नराम्रो, महँगो, पातलो, सानो, डर, फोहोर, भारी, कठोर, ढिलासुस्ती, असुहाउँद, दिलो, असफल, अन्याय, असक्षम, गरिब

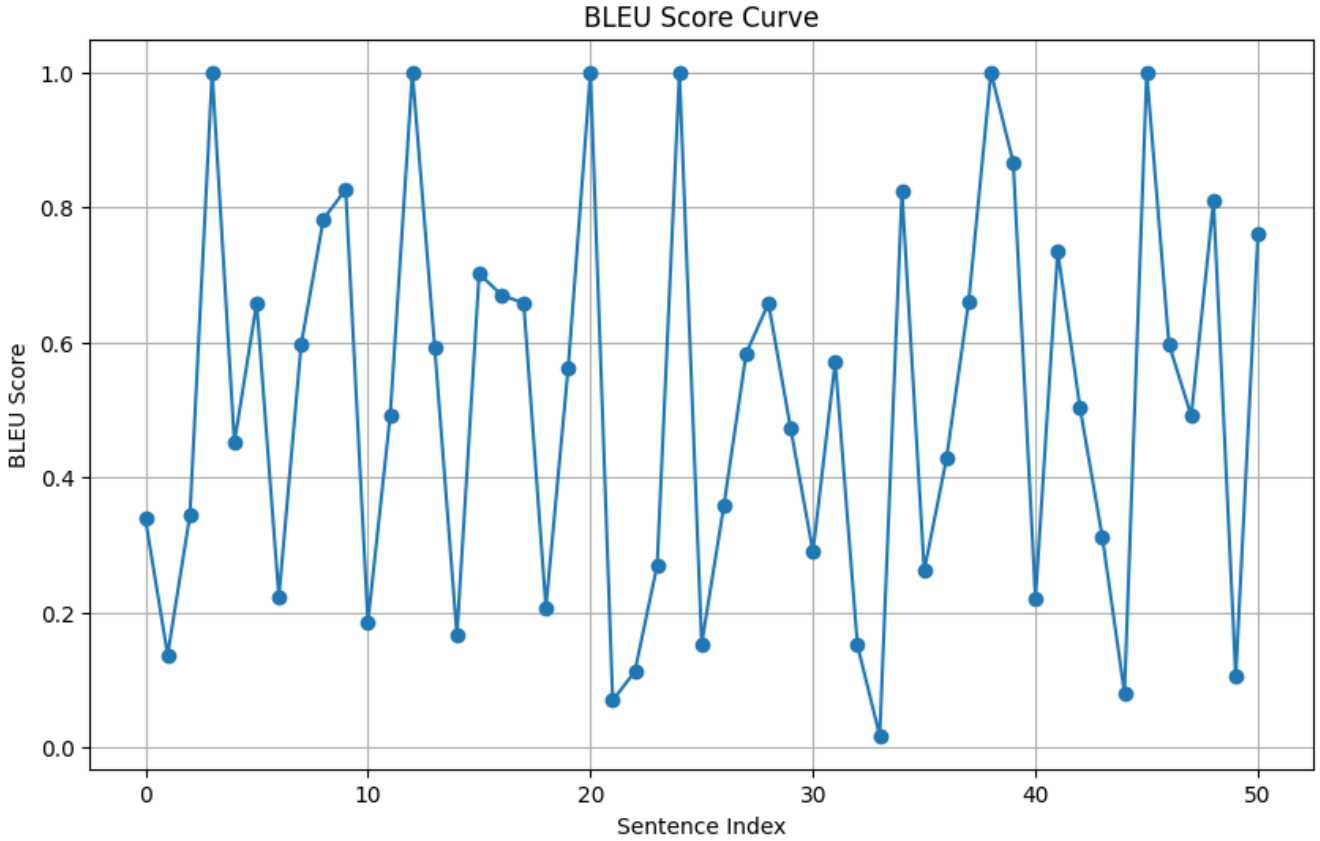
Discussion-[3]
(Clusters Visualization)

Discussion-[4]
(Test Set)

Original Sentence
पानी कम हुने सिजनमा बाँध भत्किएको भन्दा पनि भत्काइएको हुन सक्ने उनी दाबी गर्छन्
भन्नुहुन्छ रुचि त छ तर राम्रो कथा र निर्माता भेटेमा
गाईवस्तु सकुशल भेटेमा र सञ्चो भएमा परेवाको बली दिइन्छ
सबै ठाउँमा हामीले आफूलाई लागेका कुराहरु बोल्यौं
गाउँले खोल्साखोल्सीमै सौच गर्थे

Incorrect Sentence
पानी कम हुने सिजनमा बाँध भत्किएको भन्दा पनि भत्काइएको हुन सक्ने उनी दाबी गर्जिन्
भन्नुहुन्छ रुचि त छ तर राम्रो कथा र निर्माता भेगमा
गाईवस्तु सकुशल भलीमा र सञ्चो भएमा परेवाको बली दिइन्छ
सबै ठाउँमा हामीले आफूलाई लागेका कुराहरु बोलेनौं
गाउँले खोल्साखोल्सीमै साचि गर्थे

Discussion-[5]



Average BLEU Score: 50.67

Remaining Tasks

- To use Tranformer based language model.
- Integrate NER to correctly identify and handle named entities which may not follow standard spelling rules.
- Develop an interactive user interface that provides real-time suggestions and allows users to choose from multiple correction options.
- Expand vocabulary.

References-[1]

[1] A. M. Turing, Computing machinery and intelligence. Springer, 2009.

[2] P. Gupta, “A context-sensitive real-time spell checker with language adaptability,” 2020, 10.1109/ICSC.2020.00023. [Online]. Available: 10.1109/ICSC.2020.00023

[3] B. Prasain, N. lamichhane, N. Pandey, P. Adhikari, and P. Mudbhari, “Nepali spell checker,” 2023, https://doi.org/10.3126/jes2.v1i1.58461.

[4] S. Bista, Kumar, B. Keshari, L. Khatiwada, Prasad, P. Chitrakar, and S. Gurung, “Nepali lexicon development,” 2004-2007, https://www.yumpu.com/en/document/view/25135568/nepali-lexicon-development-pan-localization.

[5] X. Ziang, A. Anand, A. Naveen, J. Dan, and A. Y. Ng, “Neural language correction with character-based attention,” 2016, https://doi.org/10.48550/arXiv. 1603.09727. [6] N. Luitel, N. Bekoju, A. Kumar Sah, and S. Shakya, “Contextual spelling correction with language model for low-resource setting,” 2024, https://doi.org/10.48550/arXiv.1603.09727.

[7] A. PAL1 and A. MUSTAFI2, “Automatic context-sensitive spelling correction of ocr-generated hindi text using bert and levenshtein distance,” 2020, https://doi.org/10.48550/arXiv.2012.076527.

References-[2]

[8] Y. Bassil and M. Alwani, “A context-sensitive spelling correction using google web 1t 5-gram information,” 2020, https://doi.org/10.48550/arXiv.1204.5852.

[9] B. Rijal and S. B. Basnet, “Vector distance based spelling checking systemin nepali with language-dependent,” 2020.

[10] B. Rijal, S. Basnet, S. Awale, and S. Prasai, “Preprocessing of nepali news corpus for downstream tasks,” 2022, https://doi.org/10.3126/nl.v35i01.46553.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” CoRR, vol. abs/1409.3215, 2014. [Online]. Available: http://arxiv.org/abs/1409.3215 47

[12] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.