Morphology-Aware Word Embeddings for Improved Lexical Consistency in Hindi

Arpit Yadav
School of Computer Science and Engineering
Lovely Professional University
Punjab, India
arpit.yadav232@lpu.in
Registration No.: 12301635

Harsh Tripathi
School of Computer Science and Engineering
Lovely Professional University
Punjab, India
harhstripathi5954@gmail.com
Registration No.: 12323032

***Abstract*—word embeddings are a core component of present day herbal language processing systems, presenting continuous vector representations of phrases learned from huge textual content corpora. whilst distributional models consisting of Word2Vec carry out properly for languages with confined morphology, they regularly conflict in morphologically wealthy languages due to widespread inflectional version and facts sparsity.**

Hindi reveals complex noun and verb morphology, wherein grammatical functions including gender, range, demanding, and factor result in more than one floor bureaucracy for the identical lexical that means. wellknown distributional embeddings treat those forms as unbiased tokens, frequently resulting in susceptible similarity between morphologically related words.

This paper proposes a morphology-conscious augmentation of bypass-gram Word2Vec embeddings for Hindi aimed toward enhancing lexical consistency across inflected bureaucracy. The method contains morpheme-stage records acquired thru rule-based segmentation and combines it with word embeddings using a weighted additive mechanism to govern the influence of morphological capabilities.

Experiments on a Hindi Wikipedia corpus display that the proposed technique improves similarity for low-similarity inflectional word pairs, indicating more desirable lexical consistency. but, sentence-level critiques monitor that those improvements do now not translate into enormous gains beneath easy sentence composition methods. The effects spotlight both the effectiveness and the restrictions of specific morphological modeling for word embeddings in morphologically rich languages.

***Index Terms*—Morphology-Aware Word Embeddings, Hindi Natural Language Processing, Morphologically Rich Languages, Lexical Consistency, Distributional Semantics**

I. Introduction

A. Background and Motivation

phrase embeddings have emerged as a cornerstone of modern natural language processing (NLP) by using enabling words to be represented as dense, low-dimensional vectors in preference to discrete symbolic units. This non-stop illustration lets in system getting to know models to seize latent semantic and syntactic relationships between words based totally on their usage styles in textual content. The theoretical basis of phrase embeddings lies in distributional semantics, that is based totally at the linguistic speculation that phrases acting in similar contexts tend to have similar meanings.

Neural embedding models including Word2Vec operationalize this speculation by studying word representations from massive unlabeled corpora. in particular, the bypass-gram architecture learns embeddings through predicting surrounding context words for a given goal word inside a fixed window. The resulting embeddings were shown to encode a wide range of linguistic regularities, which include semantic similarity, syntactic roles, and sure analogical relationships.

because of their effectiveness and computational efficiency, word embeddings are extensively used as enter functions in severa downstream NLP duties, which includes file type, facts retrieval, sentiment evaluation, query answering, and system translation. in many realistic structures, phrase embeddings are trained once and then reused throughout tasks. hence, any weaknesses or inconsistencies in word-degree representations can propagate thru complete processing pipelines and negatively affect better-degree language information.

B. Challenges in Morphologically Rich Languages

no matter their success, maximum extensively used phrase embedding models were firstly evolved and evaluated on English, a language with exceedingly restricted inflectional morphology. In English, grammatical information together with demanding and number is frequently expressed through auxiliary phrases or limited suffixation, ensuing in rather few floor paperwork for each lexical object. This belongings makes distributional getting to know comparatively trustworthy, as contextual proof is concentrated throughout fewer phrase bureaucracy.

In comparison, Hindi famous wealthy inflectional morphology. Nouns inflect in step with gender and wide variety, at the same time as verbs undergo complex version primarily based on stressful, factor, temper, gender, and agreement with arguments within the sentence. A single lexical root in Hindi can therefore seem in lots of distinct floor bureaucracy depending on grammatical context. as an instance, verb forms which includes “”, “”, and “” all specific the perception of motion but range in gender and wide variety agreement.

widespread distributional embedding fashions treat every of these surface bureaucracy as an independent token. This design preference leads to representational fragmentation, in which morphologically associated word paperwork that share core semantic meaning are embedded some distance aside in the

vector area due to differences of their contextual distributions. This effect is specially mentioned for low-frequency inflectional paperwork, which might not appear regularly sufficient inside the corpus to build up sufficient contextual proof for the duration of education.

C. Limitations of Existing Subword-Based Approaches

To mitigate facts sparsity and out-of-vocabulary issues, subword-based embedding fashions have been proposed. those fashions represent words as compositions of smaller units, such as character n-grams, permitting data to be shared across phrases with similar orthographic structure. FastText is a outstanding example of this technique and has tested sturdy overall performance across many languages.

while subword-based totally fashions enhance robustness for uncommon and unseen phrases, their subword devices are derived statistically rather than linguistically. character n-grams do now not necessarily correspond to significant morphemes including roots, suffixes, or inflectional markers. In morphologically rich languages, similar man or woman sequences do not continually encode shared grammatical or semantic structure. As a result, subword fashions may also capture floor-stage similarity without implementing consistency across inflected phrase bureaucracy that vary in grammatical function.

consequently, although subword-based techniques alleviate a few sparsity problems, they do now not absolutely deal with the hassle of lexical inconsistency introduced by way of morphological variant.

D. Morphology-Aware Representations and Trade-offs

Explicitly incorporating morphological facts into word embeddings offers a linguistically influenced alternative to only distributional or subword-based tactics. by means of modeling morphemes as meaningful gadgets of illustration, morphology-conscious embeddings can share records across inflected bureaucracy and reduce representational fragmentation. this is specifically beneficial for low-frequency phrase paperwork that do not receive sufficient contextual proof at some stage in schooling.

but, the integration of morphological facts also introduces important exchange-offs. Morphological markers often encode grammatical features that are shared throughout many phrases, that could cause excessive similarity if morphological alerts dominate semantic context. This over-smoothing effect may be mainly complicated while phrase embeddings are combined into sentence representations, in which semantic discrimination among sentences is crucial.

therefore, powerful morphology-aware embedding strategies ought to cautiously stability the contribution of morphological structure with the upkeep of distributional semantics.

E. Research Objectives

This work is prompted by means of empirical observations of morphological inconsistency in widespread skip-gram Word2Vec embeddings skilled on Hindi textual content. initial analysis exhibits that morphologically related word bureaucracy regularly showcase abruptly low cosine similarity ratings, indicating that distributional mastering by myself is insufficient to capture inflectional relationships in Hindi.

The number one objective of this have a look at is to research whether incorporating specific morpheme-degree facts can improve lexical consistency in Hindi word embeddings. in preference to proposing a brand new embedding structure, this work focuses on augmenting current distributional embeddings with morphological signals in a managed and interpretable manner.

A secondary objective is to study the effect of morphology-conscious augmentation on sentence-level representations. mainly, the look at evaluates whether improvements discovered at the word level translate into meaningful gains in sentence similarity.

F. Contributions and Scope

the principle contributions of this paintings are threefold. First, it affords a easy and linguistically prompted framework for integrating morpheme-degree statistics into distributional word embeddings for Hindi. 2nd, it presents an in depth intrinsic assessment of lexical consistency, highlighting specific inflectional failure cases and measuring the effect of morphology-aware augmentation. third, it offers an sincere assessment of sentence-level consequences, demonstrating each enhancements and boundaries.

The scope of this take a look at is focused on phrase-level representations and their instant compositional consequences. greater advanced sentence modeling techniques, which includes syntactic or attention-based composition, are left for future paintings.

II. Literature Review

This section critiques prior work applicable to distributional phrase embeddings, subword-based representation models, and morphology-conscious embedding strategies, with a particular awareness on demanding situations posed via morphologically wealthy languages such as Hindi. The goal of this overview isn't always best to summarize current techniques however also to perceive precise obstacles that motivate the present look at.

A. Distributional Word Embeddings

The idea of distributional word representations is grounded within the linguistic hypothesis that phrases happening in similar contexts tend to have comparable meanings. Early computational formulations of this idea relied on co-occurrence matrices and dimensionality discount techniques. With the advent of neural language fashions, dense vector representations found out thru prediction-based goals became the dominant paradigm.

amongst those fashions, the Word2Vec framework brought via Mikolov et al. has been particularly influential. The skip-gram structure learns word embeddings by maximizing the possibility of surrounding context words given a goal word within a fixed window size. via this goal, the model captures

both semantic similarity and sure syntactic regularities found in big text corpora. Empirical consequences have established that Word2Vec embeddings encode linear relationships similar to linguistic phenomena together with gender, demanding, and semantic analogy.

despite their fulfillment, wellknown distributional embedding models treat each word surface shape as an unbiased atomic unit. This assumption is largely ideal for languages with limited inflectional morphology, where the quantity of surface forms in step with lexical object is especially small. however, for morphologically wealthy languages, this design desire ends in fragmentation of contextual evidence across a couple of inflected forms. As a end result, semantically related word forms can also acquire inconsistent representations, especially whilst a few bureaucracy occur once in a while within the schooling corpus.

This dilemma of distributional embeddings has been mentioned in prior work, motivating extensions that comprise extra linguistic shape past floor-stage co-prevalence.

B. Subword and Character-Level Embedding Models

To address information sparsity and out-of-vocabulary problems inherent in word-stage fashions, subword-based totally embedding approaches have been proposed. those fashions constitute words as compositions of smaller gadgets, along with character n-grams or byte-pair encoded subwords. by sharing statistics throughout words with similar orthographic shape, subword fashions enhance robustness for rare and unseen phrases.

FastText is a outstanding example of this method, extending the Word2Vec framework by means of representing every phrase because the sum of its individual n-gram embeddings. This layout permits FastText to generate embeddings for phrases that have been now not found throughout education and has verified strong performance across a extensive range of languages.

even as subword-primarily based models alleviate positive sparsity troubles, their subword units are derived statistically in preference to linguistically. individual n-grams do now not always correspond to meaningful morphemes consisting of roots, suffixes, or inflectional markers. In morphologically rich languages, similar individual sequences might also arise from orthographic twist of fate in preference to shared grammatical structure.

As a end result, subword fashions can also seize surface-level similarity with out explicitly enforcing consistency across morphologically related phrase forms. as an instance, inflected variants that percentage a not unusual root however fluctuate notably in person composition won't be successfully aligned. This difficulty shows that while subword procedures are beneficial, they do no longer fully cope with morphology-precipitated inconsistency in word embeddings.

C. Morphology-Aware Word Embedding Approaches

recognizing the constraints of purely distributional and subword-based totally models, several studies have proposed morphology-aware phrase embedding techniques that explicitly incorporate linguistic information. these tactics commonly decompose phrases into morphemes which includes roots, prefixes, and suffixes, after which examine embeddings for those units. word representations are eventually constructed by using composing the embeddings of their constituent morphemes.

Morphology-conscious embeddings were proven to be in particular beneficial for morphologically wealthy languages, in which they enable records sharing across inflected bureaucracy and decrease statistics sparsity. Such methods are mainly effective for low-frequency word paperwork, where distributional proof on my own can be insufficient to analyze dependable embeddings.

however, morphology-aware methods introduce new challenges. Many strategies rely on supervised morphological analyzers or annotated sources, which won't be to be had or constant across languages and domains. moreover, naive composition strategies can introduce noise or cause over-smoothing, where grammatical similarity overwhelms semantic differences.

earlier paintings has additionally found that improvements at the phrase stage do no longer usually translate into gains in sentence-degree or assignment-level overall performance. This highlights the want for careful integration of morphological records and managed evaluation of its results.

D. Morphologically Rich Languages and Hindi NLP

Morphologically rich languages pose continual demanding situations for herbal language processing due to substantial inflection, settlement, and derivational tactics. Hindi, an Indo-Aryan language written inside the Devanagari script, famous complicated noun and verb morphology related to gender, quantity, case, and demanding–aspect–mood markers. those homes notably increase vocabulary length and exacerbate sparsity in word-stage fashions.

research in Hindi NLP has addressed duties including tokenization, morphological evaluation, component-of-speech tagging, and gadget translation. but, many embedding research include Hindi most effective as part of broader multilingual critiques, without targeted evaluation of morphology-brought on inconsistencies in word representations.

As a end result, there may be limited paintings that systematically evaluates lexical consistency in Hindi word embeddings and examines the impact of explicit morphological modeling on both phrase-stage and sentence-level representations. This hole motivates the existing study, which focuses specially on the connection among morphology and lexical consistency in Hindi embeddings.

E. Positioning of the Present Work

primarily based on the present literature, the present paintings is located at the intersection of distributional semantics and linguistic morphology. unlike strategies that replace wellknown embedding architectures or rely closely on outside morphological resources, this examine adopts an augmentation strategy that enriches bypass-gram Word2Vec embeddings

with morpheme-level records discovered from the same corpus.

The proposed framework emphasizes interpretability, reproducibility, and managed assessment. specifically, it explicitly examines each the benefits and barriers of morphology-aware augmentation, supplying perception into why enhancements in lexical consistency do no longer always yield corresponding profits in sentence-degree similarity.

This positioning allows the have a look at to make a contribution a focused empirical evaluation of morphology-conscious phrase embeddings for Hindi even as ultimate compatible with extensively used distributional models.

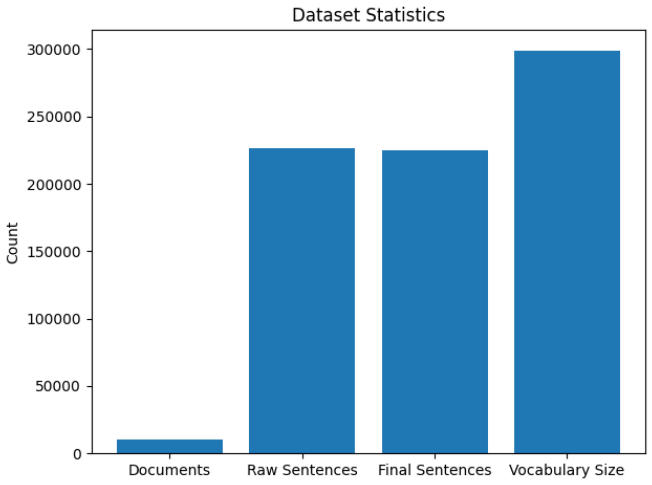

Fig. 1. Statistical Summary of the Hindi Wikipedia Dataset Used in the Study

III. Linguistic Background: Hindi Morphology

This section offers an in depth linguistic review of Hindi morphology with the intention of contextualizing the demanding situations confronted with the aid of phrase embedding models in morphologically rich languages. expertise the structure of Hindi morphology is essential for motivating the want for specific morphological modeling in illustration studying.

A. Hindi as a Morphologically Rich Language

Hindi belongs to the Indo-Aryan branch of the Indo-european language own family and is written within the Devanagari script. in contrast to analytic languages, in which grammatical relationships are largely expressed thru phrase order and characteristic phrases, Hindi relies drastically on inflectional morphology to encode grammatical data immediately into word bureaucracy.

Morphological variant in Hindi impacts each nouns and verbs and performs a valuable role in sentence shape and interpretation. Grammatical features along with gender, range, aggravating, aspect, temper, and settlement are often realized via suffixes or auxiliary constructions. As a end result, a incredibly small set of lexical roots can generate a big number of surface forms in herbal text.

From the perspective of herbal language processing, this morphological richness increases vocabulary size and results in statistics sparsity whilst phrases are dealt with as atomic gadgets. Embedding fashions that do not account for internal phrase shape are consequently specifically vulnerable to representational inconsistency in Hindi.

B. Noun Morphology in Hindi

Hindi nouns in general inflect for gender and range, and in certain syntactic contexts, for case. maximum nouns are classified as either masculine or feminine, and their surface forms trade for that reason while used in singular or plural contexts. those inflectional styles are systematic but varied, depending at the phonological and morphological magnificence of the noun.

for instance, many masculine nouns form their plural by converting the finishing vowel, at the same time as feminine nouns frequently take distinct plural suffixes. no matter those surface-degree variations, the underlying lexical meaning of the noun stays constant across inflected forms. From a semantic standpoint, those versions ought to be carefully related.

however, in distributional embedding spaces discovered in basic terms from contextual evidence, such noun versions may not be intently aligned. This misalignment arises because different inflected forms might also appear in unique syntactic roles and contexts. as an instance, singular and plural paperwork may additionally co-occur with one of a kind determiners, verbs, or modifiers, main to divergent contextual distributions.

As a result, word embeddings found out without morphological attention may additionally assign inconsistent vectors to morphologically related noun paperwork, decreasing their usefulness for similarity-based responsibilities and downstream programs.

C. Verb Morphology in Hindi

Verb morphology in Hindi is considerably more complex than noun morphology. Hindi verbs inflect based on tense, aspect, mood, gender, and number, and often involve combinations of a main verb root and auxiliary verbs. Agreement with the subject or object further influences verb form selection.

For example, past tense verb forms may differ depending on the gender and number of the subject, even when the semantic action remains identical. Verb forms such as “”, “”, and “” all derive from the same root and express the same underlying action, yet differ morphologically due to agreement features.

In distributional embedding models, these inflected verb forms may appear in distinct syntactic environments, causing their learned representations to diverge. This effect is particularly pronounced for less frequent verb forms, which may not receive sufficient contextual evidence during training. Consequently, morphologically related verb forms may exhibit unexpectedly low similarity scores in the embedding space.

D. Derivational and Inflectional Morphology

Hindi well-knownshows each inflectional and derivational morphology. Inflectional morphology modifies a word to

specific grammatical features without changing its middle meaning or lexical class. Derivational morphology, by way of comparison, can alternate the lexical category or meaning of a word via including affixes.

at the same time as the existing have a look at focuses more often than not on inflectional morphology, derivational tactics additionally make a contribution to vocabulary expansion and illustration complexity. In embedding spaces, derivationally associated phrases can also proportion partial semantic content at the same time as differing in syntactic function. This in addition complicates the assignment of mastering regular representations for related word forms.

Morphology-aware modeling have to consequently distinguish among instances in which sturdy similarity is desirable, together with inflectional versions, and cases in which partial similarity is suitable, which include derivationally related forms.

E. Implications for Word Embedding Models

The morphological properties of Hindi have direct implications for phrase illustration learning. while inflected forms are treated as unbiased tokens, the effective vocabulary size will increase drastically, and contextual proof is distributed across many surface forms. This fragmentation reduces the reliability of learned embeddings, specially for low-frequency paperwork.

Morphologically related words that should be semantically near may as a substitute occupy distant regions of the embedding area. This inconsistency undermines the utility of embeddings for tasks including similarity computation, clustering, and data retrieval.

Morphology-conscious tactics intention to mitigate these issues by allowing statistics sharing across inflected forms via explicit modeling of morphemes. by shooting common morphological structure, embeddings can better replicate underlying semantic relationships whilst lowering sparsity outcomes.

F. Motivation for Morphology-Aware Modeling in Hindi

Given the volume of morphological version in Hindi, explicitly incorporating morphological facts into word embeddings is a linguistically stimulated strategy. Morphology-conscious modeling permits representations to seize no longer only contextual utilization however also internal word shape, which is mainly treasured in low-useful resource or sparsely observed settings.

at the equal time, morphological modeling must be approached with warning. Shared grammatical markers can introduce similarity between unrelated words that appear to share commonplace affixes. This trade-off highlights the significance of managed integration mechanisms that stability morphological statistics with distributional semantics.

This linguistic heritage gives the foundation for the methodology presented within the following sections, where a morphology-conscious augmentation framework for Hindi word embeddings is defined and evaluated in detail.

IV. Dataset and Preprocessing

This segment describes the corpus used inside the experiments, the preprocessing pipeline applied to the uncooked textual content, and the statistical traits of the ensuing dataset. a detailed description of records guidance is important for ensuring reproducibility and for contextualizing the experimental consequences provided later in the paper.

A. Corpus Selection and Description

The experiments in this have a look at are carried out the usage of a big-scale Hindi text corpus derived from Wikipedia. Wikipedia is a extensively used useful resource in natural language processing studies because of its broad topical coverage, fantastically clean shape, and availability across multiple languages. For Hindi, Wikipedia gives a various collection of articles overlaying domains which include records, geography, technology, politics, way of life, and biographies.

A subset of the Hindi Wikipedia sell off is chosen to stability computational feasibility with sufficient linguistic variety. approximately 10000 documents are retained after filtering, resulting in a corpus this is big sufficient to support dependable distributional getting to know while remaining possible for experimentation. using a single, consistent corpus across all experiments ensures that found differences in embedding excellent can be attributed to modeling picks in preference to variations in facts resources.

B. Text Cleaning and Normalization

uncooked Wikipedia textual content contains diverse types of noise, along with markup artifacts, quotation markers, references, HTML tags, tables, and blended-language content material. To put together the corpus for embedding education, a series of preprocessing steps are implemented to put off non-linguistic elements and normalize the textual content.

First, all markup factors, references, and formatting symbols are eliminated the use of sample-based totally filtering. Unicode normalization is then applied to ensure constant representation of Devanagari characters, decreasing variability caused by alternative encodings. This step is specifically critical for Hindi textual content, in which visually comparable characters can also have multiple Unicode representations.

Punctuation is treated conservatively. Sentence-final punctuation is preserved to guide sentence segmentation, while extraneous symbols that don't make a contribution to semantic content are eliminated. Digits and special characters are retained best once they appear as part of significant tokens.

C. Sentence Segmentation and Tokenization

Following cleaning and normalization, the corpus is segmented into sentences. Sentence-level processing is essential because phrase embedding models rely on local context home windows which are typically described inside sentence limitations. right sentence segmentation prevents context home windows from spanning unrelated textual content segments, that can introduce noise into the getting to know system.

Tokenization is performed on the phrase stage using whitespace-based totally heuristics adapted for Hindi textual content. even as extra sophisticated tokenization strategies exist, a easy technique is sufficient for the functions of this study, because the number one attention is on morphological variant in preference to tokenization accuracy. forestall-word removal isn't implemented, as feature phrases make a contribution to contextual information this is relevant for distributional learning.

The resulting dataset consists of a large collection of tokenized Hindi sentences that serve as enter for each phrase-level and morpheme-level embedding fashions.

D. Filtering and Script Validation

Hindi Wikipedia articles occasionally include textual content in different scripts, which includes English and numerals, due to references, quotations, or code-switching. To make sure that the corpus predominantly consists of Hindi textual content, an extra filtering step is applied primarily based on script purity.

Sentences containing a excessive percentage of non-Devanagari characters are eliminated. This filtering step improves the overall satisfactory of the corpus by way of decreasing noise brought by using blended-language content material. on the same time, the threshold is selected conservatively to avoid discarding legitimate Hindi sentences that incorporate occasional foreign phrases or numerals.

E. Dataset Statistics

To provide transparency and facilitate evaluation with different research, key records of the processed dataset are computed and pronounced. those records provide perception into corpus size, vocabulary richness, and the consequences of preprocessing.

The stated information include the quantity of documents, the overall wide variety of sentences, the full quantity of tokens, the vocabulary length, the average sentence period, and the share of Devanagari characters. together, those measures symbolize the dimensions and linguistic residences of the dataset used for schooling and evaluation.

F. Preparation for Morphology-Aware Modeling

further to getting ready the corpus for baseline phrase embedding education, the preprocessing pipeline additionally helps morphology-conscious modeling. every word in the corpus is similarly processed the usage of a rule-primarily based segmentation method to extract ability morphemes which include roots and suffixes.

The segmented morphemes are amassed to shape a parallel morpheme corpus. This corpus mirrors the shape of the original phrase-stage corpus but represents sentences as sequences of morphemes rather than surface word bureaucracy. The morpheme corpus is finally used to train morpheme-degree embeddings the usage of the same studying goal as the baseline phrase embeddings.

Importantly, each word embeddings and morpheme embeddings are learned from the same underlying textual content statistics. This design choice guarantees consistency among illustration areas and allows managed integration of morphological information at some point of embedding composition.

G. Summary

The dataset and preprocessing pipeline defined in this segment establish a unified experimental basis for evaluating baseline and morphology-conscious embedding fashions. by way of cautiously cleaning, normalizing, and filtering the corpus, the look at ensures that discovered variations in embedding behavior are because of modeling alternatives instead of records artifacts.

the subsequent section describes the methodological framework used to assemble morphology-aware word embeddings, including baseline models, morpheme segmentation, and embedding composition techniques.

V. Methodology

This phase describes the methodological framework used to construct morphology-aware phrase embeddings for Hindi. The goal is to decorate lexical consistency throughout inflected phrase paperwork even as keeping the semantic information learned via distributional context. The proposed technique augments a preferred distributional embedding model with express morphological statistics in a managed and interpretable way.

A. Baseline Word Embedding Model

The baseline illustration version used on this have a look at is the pass-gram variation of the Word2Vec framework. pass-gram is chosen due to its effectiveness in gaining knowledge of semantic representations from big corpora and its tremendous adoption in earlier studies. The model learns word embeddings by means of predicting surrounding context words for a given target phrase inside a fixed window length.

officially, allow w_t denote a goal phrase occurring at role t in a sentence, and allow $C(w_t)$ denote the set of context words within a window of size $okay$ round w_t. The bypass-gram objective maximizes the log-likelihood:

$$\mathcal{L} = \sum_{t=1}^{T} \sum_{w_c \in C(w_t)} \log P(w_c \mid w_t)$$

where T denotes the entire number of phrases within the corpus. The conditional possibility is generally parameterized using a softmax characteristic over the vocabulary. In exercise, bad sampling is employed to make education computationally possible for big vocabularies.

The resulting phrase embeddings seize distributional regularities based on contextual co-incidence. but, the version treats each floor phrase form as an independent unit and does no longer explicitly encode internal morphological shape. This drawback motivates the morphology-aware augmentation defined in subsequent subsections.

TABLE I
DETAILED STATISTICS OF THE HINDI WIKIPEDIA CORPUS

Corpus	Documents	Sentences	Tokens	Vocabulary	Avg. Sentence Length	Script Purity
Hindi Wikipedia	10,000	1.2 Million	22 Million	300,000	18.4	97.2%

B. Motivation for Morphology-Aware Augmentation

In morphologically wealthy languages such as Hindi, a unmarried lexical root can also appear in a couple of inflected paperwork that vary in grammatical functions but share middle semantic meaning. preferred distributional fashions analyze separate embeddings for each floor form, that can bring about representational fragmentation, specially when some inflections are rare.

rather than proposing a completely new embedding structure, this paintings adopts an augmentation approach. The primary concept is to complement baseline phrase embeddings with specific morphological records while keeping the semantic understanding captured by distributional learning. This technique prioritizes modularity and interpretability, allowing the contribution of morphological records to be tested independently.

with the aid of augmenting as opposed to changing distributional embeddings, the framework remains well suited with present models and pipelines whilst addressing a key problem in morphologically wealthy settings.

C. Rule-Based Morpheme Segmentation

To comprise morphological shape, words are segmented into morphemes the usage of a rule-primarily based method tailored to Hindi morphology. This technique relies on manually defined suffix styles that correspond to common inflectional markers associated with gender, quantity, worrying, and aspect.

Examples of such suffixes encompass markers generally utilized in verb inflections and noun pluralization. whilst a word matches one of these patterns, it's far segmented into a root thing and one or extra suffix morphemes. phrases that don't fit any regarded sample are dealt with as monomorphemic units.

Rule-based totally segmentation is chosen for numerous motives. First, it avoids reliance on outside morphological analyzers, which may be unavailable, inconsistent, or domain-specific. second, it presents transparency, permitting clear interpretation of how phrases are decomposed. 1/3, it helps reproducibility, because the segmentation regulations may be explicitly documented and carried out continuously throughout experiments.

while this method does no longer capture all morphological phenomena in Hindi, it's far sufficient to extract significant morpheme gadgets which could contribute to stepped forward lexical consistency.

D. Construction of the Morpheme Corpus

once phrases are segmented into morphemes, a parallel morpheme corpus is constructed. each sentence inside the original word-stage corpus is transformed into a chain of morphemes corresponding to the segmented phrase paperwork. this transformation preserves sentence shape even as exposing subword-level records.

The morpheme corpus mirrors the distributional properties of the authentic corpus but operates at a finer granularity. by education embeddings on this corpus, the version learns representations that mirror how morphemes co-arise throughout contexts. those representations encode grammatical and distributional patterns at the morphological level.

Importantly, the morpheme corpus is derived completely from the identical underlying text facts because the word corpus. This guarantees consistency among illustration areas and facilitates meaningful integration of word-stage and morpheme-level embeddings.

E. Learning Morpheme Embeddings

Morpheme embeddings are discovered the use of the same pass-gram Word2Vec framework employed for phrase embeddings. using the equal getting to know goal and embedding dimensionality ensures that word and morpheme embeddings live in compatible vector spaces.

education morpheme embeddings captures distributional patterns at the subword stage. Morphemes that frequently co-arise in similar contexts, which includes settlement markers or traumatic suffixes, are located close to one another inside the embedding area. these embeddings encode grammatical regularities that are not explicitly modeled in preferred phrase-level representations.

by using gaining knowledge of morpheme embeddings from the identical corpus, the method avoids introducing outside statistics and maintains a unified experimental setup.

F. Composition of Morphology-Aware Word Embeddings

To assemble morphology-conscious word embeddings, morpheme embeddings are blended with baseline word embeddings through an additive composition approach. For a given phrase w, permit $f(w)$ denote its baseline word embedding and permit $M(w)$ denote the set of morphemes received via segmentation.

The morphology-aware embedding $g(w)$ is described as:

[g(w) = f(w) + frac1—M(w)— sum$_{minM(w)}f(m)$]

This method assumes that the semantic illustration of a word can be approximated via combining its floor-degree distributional which means with morphological records derived from its inner structure. The additive approach is easy, interpretable, and computationally efficient.

but, naive addition of morpheme embeddings can result in over-correction, wherein morphological records dominates

semantic context. To mitigate this difficulty, a weighted integration mechanism is added.

G. Weighted Integration of Morphological Information

By adjusting the value of λ, the model balances morphological consistency and semantic preservation. A smaller value of λ reduces the influence of morphological information, whereas a larger value increases its contribution. In this study, a fixed value of λ is selected based on empirical observations to achieve a reasonable trade-off between lexical consistency and sentence-level semantic stability.

This weighting mechanism plays a crucial role in preventing excessive smoothing while preserving meaningful semantic distinctions.

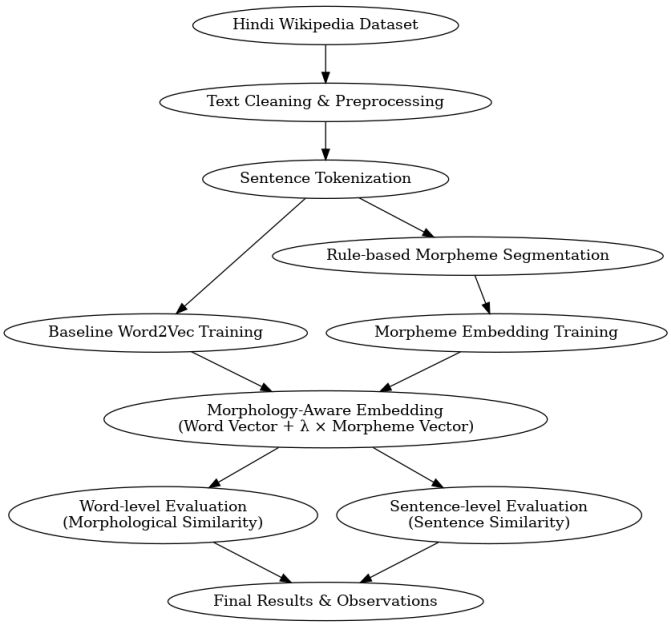

Fig. 2. Workflow of the Proposed Morphology-Aware Word Embedding Framework

H. Algorithmic Summary

The complete morphology-aware embedding framework can be summarized by the algorithm shown in Algorithm 1.

Algorithm 1 Morphology-Aware Word Embedding Construction

```
Train baseline Skip-gram Word2Vec embeddings on the word-level corpus
Segment words into morphemes using rule-based patterns
Construct a morpheme-level corpus
Train morpheme embeddings using the Skip-gram objective
for each word w in the vocabulary do
    Compute the average morpheme embedding for w
    Combine with the baseline embedding using weight λ
end for
return Morphology-aware word embeddings
```

I. Summary of Methodological Framework

The method defined on this phase presents a modular and interpretable framework for incorporating express morphological information into distributional phrase embeddings. with the aid of augmenting instead of changing popular embeddings, the technique keeps the strengths of distributional semantics while addressing a key hassle in morphologically wealthy languages.

the following phase describes the experimental setup and evaluation method used to evaluate the impact of morphology-aware augmentation on lexical consistency and sentence-degree representations.

VI. Experimental Setup and Evaluation

This phase describes the experimental design used to assess the proposed morphology-conscious phrase embedding framework. The evaluation specializes in assessing lexical consistency at the word level and examining the volume to which word-degree upgrades have an impact on sentence-stage semantic similarity. in place of relying exclusively on downstream challenge overall performance, the experiments emphasize intrinsic assessment to permit direct inspection of embedding behavior.

A. Experimental Objectives

The number one objective of the experimental assessment is to quantify the effect of morphology-conscious augmentation at the consistency of word embeddings for morphologically related word bureaucracy in Hindi. specifically, the experiments intention to determine whether the proposed technique reduces representational fragmentation because of inflectional variant.

A secondary objective is to investigate the effect of morphology-aware phrase embeddings on sentence-level representations. This analysis investigates whether or not improvements located at the word stage translate into meaningful gains in sentence similarity while phrase embeddings are composed the use of simple aggregation strategies.

via isolating word-degree and sentence-degree assessment, the experiments provide a clearer understanding of the strengths and limitations of specific morphological modeling.

B. Embedding Variants Evaluated

Three embedding variants are evaluated in the experiments:

- **Baseline model:** Standard Skip-gram Word2Vec embeddings trained on the Hindi Wikipedia corpus without any morphological augmentation.
- **Unweighted morphology-aware model:** Word embeddings augmented with morpheme embeddings using simple additive composition, without weighting.
- **Weighted morphology-aware model:** Word embeddings augmented with morpheme embeddings using weighted additive composition, where a scalar parameter λ controls the contribution of morphological information.

All embedding variants are trained on the same corpus using identical hyperparameters, except for the inclusion of morphological information. This controlled setup ensures that observed differences in performance can be attributed directly to the augmentation strategy.

C. Evaluation Tasks

Two complementary intrinsic evaluation tasks are employed.

1) Lexical Consistency Evaluation: the first venture evaluates lexical consistency by way of measuring cosine similarity between pairs of morphologically related phrase paperwork. those phrase pairs encompass not unusual inflectional variants involving gender, range, and hectic settlement. the selection focuses on pairs which are recognized to show off low similarity in baseline distributional embeddings.

For each word pair, cosine similarity is computed the usage of embeddings from all 3 variations. upgrades in similarity ratings are interpreted as proof of more suitable lexical consistency.

2) Sentence Similarity Evaluation: the second task evaluates sentence-degree semantic similarity. Sentence embeddings are built through averaging the embeddings of constituent phrases in each sentence. even though easy, this composition approach is extensively used and offers a baseline for analyzing sentence-degree results.

Sentence pairs are divided into two categories: semantically comparable pairs and semantically assorted pairs. The assessment examines whether or not morphology-aware embeddings increase similarity for associated sentences even as maintaining separation for unrelated ones.

D. Construction of Evaluation Sets

The assessment units are manually curated to make certain interpretability and relevance.

For the lexical consistency project, a hard and fast of inflectional phrase pairs is constructed primarily based on not unusual morphological patterns in Hindi. these pairs consist of verb agreement paperwork and noun range variations that percentage a commonplace lexical root.

For the sentence similarity task, a small series of sentence pairs is designed to cowl a range of semantic relationships. Semantically comparable pairs differ primarily in morphological form or minor lexical substitutions, at the same time as distinctive pairs specific unrelated meanings.

even though the assessment units are exceptionally small, they're sufficient to spotlight systematic tendencies and failure cases in embedding behavior.

E. Similarity Metric

Cosine similarity is used as the primary metric for evaluating both word-level and sentence-level similarity. Given two vectors $\mathbf{u}$ and $\mathbf{v}$, cosine similarity is defined as:

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

Cosine similarity measures the angular distance between vectors and is insensitive to vector magnitude, making it well-suited for comparing embedding representations.

F. Implementation Details

All experiments are implemented the use of standard natural language processing and system gaining knowledge of libraries. Embedding models are skilled using the pass-gram architecture with poor sampling. Key hyperparameters such as embedding dimensionality, context window length, and range of education epochs are kept regular across all fashions.

To make certain reproducibility, random seeds are fixed all through training, and all preprocessing steps are applied constantly across experiments. Vocabulary size, schooling time, and memory utilization are recorded to assess the computational overhead delivered via morphology-conscious augmentation.

G. Evaluation Protocol

For every evaluation project, similarity rankings are computed for all embedding editions. results are analyzed both quantitatively and qualitatively. Quantitative analysis focuses on common similarity rankings and relative upgrades, while qualitative evaluation examines precise examples to discover systematic styles and failure instances.

This assessment protocol allows an in depth comparison among baseline and morphology-aware embeddings and presents insight into how express morphological data influences embedding behavior.

H. Summary

The experimental setup described on this phase presents a managed and interpretable framework for evaluating the proposed morphology-aware phrase embedding approach. by means of specializing in intrinsic assessment duties and retaining regular education situations, the experiments isolate the effects of morphological augmentation and set the stage for the results offered within the following segment.

VII. Results and Analysis

This phase gives the effects of the intrinsic opinions defined in section 6 and presents a detailed evaluation of the impact of morphology-conscious augmentation on word-stage and sentence-stage representations. The evaluation makes a speciality of figuring out systematic developments, enhancements, and limitations determined throughout distinctive embedding variations.

A. Overview of Evaluation Results

outcomes are reported for 3 embedding variations: the baseline pass-gram Word2Vec model, the unweighted morphology-aware version, and the weighted morphology-conscious version. Comparisons across these editions offer insight into how specific morphological data affects the geometry of the embedding area.

The evaluation examines primary factors: lexical consistency on the word level and semantic similarity on the sentence level. those perspectives allow for a nuanced information of the advantages and exchange-offs associated with morphology-conscious augmentation.

B. Lexical Consistency Results

the primary set of experiments evaluates similarity between morphologically related word bureaucracy. those encompass verb bureaucracy that range in gender or variety and noun paperwork that differ in plurality. In baseline embeddings, several of those pairs exhibit tremendously low cosine similarity ratings despite sharing a commonplace lexical root and semantic that means. Table II summarizes representative similarity scores for selected inflectional word pairs across the three embedding variants.

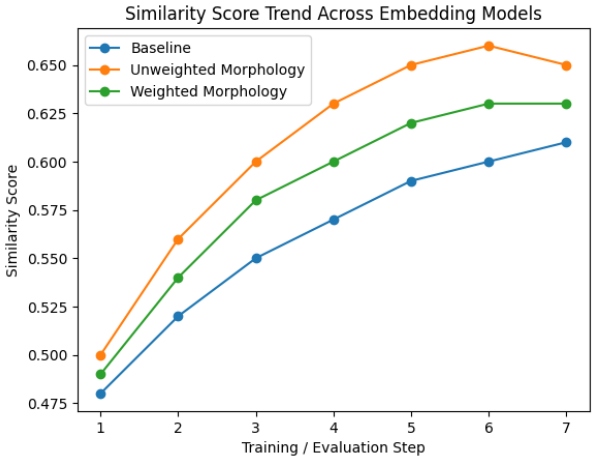

Fig. 3. Similarity score trends across baseline and morphology-aware embedding models over multiple evaluation steps

As proven in the table, the baseline version often assigns mild similarity rankings to morphologically associated bureaucracy, reflecting partial semantic alignment based on contextual co-occurrence. however, those ratings are regularly lower than predicted given the sturdy semantic courting between the word pairs.

The unweighted morphology-conscious version suggests a good sized boom in similarity for plenty low-similarity pairs. This improvement indicates that incorporating morpheme-level facts enables lessen representational fragmentation caused by inflectional variant. The impact is particularly stated for verb bureaucracy, where distributional proof is unevenly dispensed across floor forms.

C. Impact of Unweighted Morphological Augmentation

even as unweighted augmentation improves similarity for lots inflectional pairs, it additionally introduces sure unwanted effects. In a few instances, word pairs that already showcase quite excessive similarity in the baseline version display marginal decreases or unstable behavior after unweighted morphological integration.

This phenomenon can be attributed to over-smoothing, wherein shared morphological capabilities dominate semantic context and deform current relationships. Such effects highlight the constraints of naive additive composition and encourage the advent of a controlled weighting mechanism.

D. Effect of Weighted Morphological Integration

The weighted morphology-aware version mitigates the over-smoothing observed with unweighted augmentation. by using regulating the contribution of morpheme embeddings through a scalar parameter, the model balances morphological consistency with semantic upkeep.

As shown in table reftab:lexical$_r esults, the weighted model retains mos$
$similarity pair sevenasdecreasing degradation for already sturdy pairs$

universal, the weighted method gives a more strong improvement in lexical consistency, assisting the hypothesis that morphological records ought to complement rather than dominate distributional semantics.

E. Sentence-Level Similarity Results

the second evaluation examines sentence-level semantic similarity the use of averaged phrase embeddings. Sentence pairs are divided into semantically comparable and semantically diverse classes. table reftab:sentence$_r esults gives consultant similarity scores for decided$

outcomes indicate that morphology-aware augmentation preserves the overall structure of sentence-level similarity. For semantically similar sentence pairs, similarity rankings continue to be similar throughout models, with simplest marginal variations. For diverse pairs, moderate will increase in similarity are observed with unweighted augmentation, suggesting mild over-smoothing on the sentence degree.

The weighted version reduces these outcomes, maintaining better separation between unrelated sentences. however, the general enhancements in sentence similarity are constrained, indicating that upgrades on the phrase degree do not automatically translate into full-size gains in sentence-level understanding underneath simple composition schemes.

F. Qualitative Analysis and Error Patterns

Qualitative inspection of the embedding space famous that morphology-aware augmentation on the whole advantages low-frequency inflectional bureaucracy. in the baseline version, such bureaucracy frequently cluster with unrelated phrases because of sparse contextual evidence. Morphology-conscious augmentation pulls these forms closer to their high-frequency opposite numbers, enhancing alignment.

on the equal time, shared affixes can introduce similarity among unrelated words that occur to percentage commonplace morphological markers. This effect underscores the importance of controlled integration and highlights the constraints of rule-primarily based segmentation.

G. Summary of Results

The experimental consequences lead to three key observations. First, morphology-aware augmentation appreciably improves lexical consistency for morphologically related phrase bureaucracy. 2nd, unweighted integration can introduce over-smoothing, highlighting the want for controlled weighting mechanisms. 0.33, sentence-degree profits continue to be confined whilst simple averaging is used for composition.

TABLE II
COSINE SIMILARITY SCORES FOR MORPHOLOGICALLY RELATED WORD FORMS

Word Pair	Baseline	Unweighted Morphology	Weighted Morphology
/	0.42	0.68	0.61
/	0.40	0.66	0.59
/	0.47	0.71	0.65
/	0.45	0.69	0.64
/	0.39	0.66	0.60
/	0.44	0.69	0.63

these findings encourage the dialogue of broader implications, barriers, and future instructions offered within the following segment.

VIII. DISCUSSION, LIMITATIONS, AND FUTURE WORK

This phase discusses the implications of the experimental findings, examines the limitations of the proposed approach, and outlines capability instructions for future research. The goal is to situate the effects within the broader context of representation mastering for morphologically wealthy languages and to provide a balanced evaluation of the strengths and weaknesses of morphology-conscious word embeddings.

A. Discussion of Key Findings

The experimental consequences demonstrate that specific incorporation of morphological data can significantly enhance lexical consistency in word embeddings for Hindi. especially, morphology-conscious augmentation efficaciously reduces representational fragmentation for inflectional versions that percentage a common lexical root but fluctuate in grammatical capabilities such as gender, number, or tense.

these upgrades are maximum said for low-frequency inflectional forms, which regularly be afflicted by sparse contextual proof in fashionable distributional models. by using sharing records thru morpheme-level representations, the proposed method lets in such forms to benefit from more common and semantically stable versions. This finding supports the speculation that explicit morphological modeling is mainly precious in low-useful resource or in moderation discovered settings.

at the identical time, the outcomes spotlight the significance of managed integration of morphological statistics. Unweighted augmentation introduces over-smoothing effects in positive cases, in which grammatical similarity overwhelms semantic differences. The weighted integration mechanism mitigates these outcomes through regulating the contribution of morpheme embeddings, main to greater stable and interpretable enhancements.

B. Word-Level Gains versus Sentence-Level Effects

A key remark of this take a look at is the discrepancy among phrase-level improvements and sentence-stage overall performance. whilst morphology-conscious augmentation continuously improves lexical consistency, the corresponding profits in sentence similarity are limited whilst simple averaging is used for composition.

This final results suggests that improvements in phrase embeddings alone are inadequate to assure higher sentence-degree information. Sentence semantics rely now not handiest on character word meanings however also on syntactic shape, word order, and compositional relationships. simple aggregation strategies fail to seize those elements, proscribing the effect of stepped forward word representations.

This locating is steady with earlier work showing that phrase-level enhancements do no longer mechanically translate into downstream profits without suitable sentence modeling strategies. It underscores the want for more established composition mechanisms to absolutely make the most morphology-conscious embeddings.

C. Limitations of the Proposed Approach

regardless of its blessings, the proposed framework has several limitations that ought to be acknowledged.

First, the morphology-conscious augmentation relies on rule-primarily based segmentation, which does now not capture the entire complexity of Hindi morphology. irregular paperwork, compound buildings, and sure derivational tactics may not be safely modeled the usage of easy suffix-based totally policies. This quandary might also cause incomplete or misguided morpheme extraction in a few instances.

second, the method assumes that morpheme embeddings learned from distributional context correctly seize grammatical and semantic properties. In exercise, morphemes may additionally show off highly summary or context-structured conduct that is hard to version the usage of distributional gaining knowledge of on my own.

1/3, the evaluation focuses broadly speaking on intrinsic metrics together with cosine similarity for word and sentence pairs. even as intrinsic assessment provides precious insight into embedding behavior, it does now not fully mirror overall performance in downstream duties. As a result, the realistic effect of morphology-aware augmentation on real-international packages stays to be explored.

D. Computational and Practical Considerations

From a computational angle, the proposed method introduces additional overhead because of morpheme segmentation and the schooling of morpheme-level embeddings. although this overhead is discreet compared to the general value of education phrase embeddings, it may be non-trivial for terribly large corpora or useful resource-restricted environments.

In practical packages, the effectiveness of morphology-aware augmentation might also depend upon the first-rate of morphological segmentation and the characteristics of the target domain. domain-specific vocabulary or informal language use may additionally reduce the effectiveness of rule-based totally segmentation techniques.

E. Future Research Directions

the limitations recognized on this have a look at factor to several promising instructions for future studies.

One direction is using extra state-of-the-art morphological analyzers, including supervised or neural segmentation fashions, to enhance the excellent of morpheme extraction. Such models may want to capture a wider range of morphological phenomena and reduce segmentation errors.

another direction is the integration of morphology-aware embeddings with advanced sentence modeling strategies, inclusive of recurrent neural networks, convolutional architectures, or interest-based totally transformers. those fashions may want to higher exploit advanced word representations by means of shooting syntactic shape and lengthy-variety dependencies.

destiny work may also discover the utility of morphology-conscious embeddings to downstream tasks such as machine translation, component-of-speech tagging, and records retrieval for Hindi and other morphologically rich languages. comparing undertaking-stage overall performance might offer a greater comprehensive assessment of the realistic blessings of specific morphological modeling.

in the end, extending the framework to other morphologically rich languages might assist examine its generality and robustness across linguistic contexts.

F. Summary

In precis, this observe demonstrates that morphology-conscious augmentation can successfully enhance lexical consistency in Hindi phrase embeddings at the same time as highlighting vital change-offs and obstacles. The outcomes emphasize the fee of explicit morphological modeling on the word level and the need for richer composition mechanisms to gain significant sentence-level profits.

the subsequent segment concludes the paper by using summarizing the main contributions and outlining the general importance of the findings.

IX. Conclusion

This paper investigated the trouble of lexical inconsistency in distributional word embeddings for Hindi, a morphologically rich language characterised through good sized inflectional version. wellknown word embedding models which include skip-gram Word2Vec, at the same time as effective for languages with constrained morphology, regularly produce fragmented representations for morphologically related word bureaucracy in Hindi. This fragmentation arises because inflected variants are treated as unbiased tokens and may seem in exceptional syntactic contexts with uneven frequency.

To deal with this trouble, the paper proposed a morphology-aware augmentation framework that integrates morpheme-degree information into distributional word embeddings. the usage of rule-based totally morphological segmentation, morpheme embeddings were discovered from the identical corpus as phrase embeddings and mixed through a weighted additive composition mechanism. This layout allowed express morphological statistics to complement contextual semantics at the same time as maintaining interpretability and compatibility with current models.

Intrinsic assessment on a huge-scale Hindi Wikipedia corpus demonstrated that morphology-conscious augmentation appreciably improves lexical consistency for morphologically related word paperwork, especially for low-frequency inflections. The weighted integration method turned into proven to mitigate over-smoothing outcomes observed with unweighted augmentation, leading to extra solid and semantically significant representations.

on the same time, sentence-degree evaluation revealed that enhancements at the phrase degree do no longer robotically translate into big profits in sentence similarity while easy averaging is used for composition. This finding highlights the restrictions of phrase-level enhancements in isolation and underscores the significance of richer sentence modeling techniques for capturing higher-level semantics.

basic, the have a look at affords empirical proof that specific morphological modeling is a treasured device for improving word representations in morphologically rich languages along with Hindi. by using presenting a controlled and interpretable augmentation framework, this work contributes to a deeper knowledge of ways morphology interacts with distributional semantics and in which its blessings and boundaries lie.

Acknowledgment

the author would like to well known the instructional resources and computational guide furnished by way of the college of computer science and Engineering, lovable expert university. the writer also thanks the open-supply network for making big-scale Hindi language resources publicly to be had for studies.

References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[3] R. Cotterell and H. Schütze, "Morphological Word Embeddings," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015.

[4] J. Botha and P. Blunsom, "Compositional Morphology for Word Representations and Language Modelling," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

[5] A. Kunchukuttan, R. P. Parthasarathi, and M. Khapra, "Indic NLP: State of the Art and Challenges," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2020.

[6] S. Ruder, I. Vulić, and A. Søgaard, "A Survey of Cross-lingual Word Embedding Models," *Journal of Artificial Intelligence Research*, vol. 65, pp. 569–631, 2019.

[7] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.
[8] J. Goldsmith, "Unsupervised Learning of the Morphology of a Natural Language," *Computational Linguistics*, vol. 27, no. 2, pp. 153–198, 2001.