Lexical Simplification with Static Embeddings and Deep Learning A Comparative Analysis

Omar Ibrahim
C24029408
Cardiff University
ibrahimo@cardiff.ac.uk

Abstract

This study investigates two tasks of the Lexical Simplification (LS) pipeline–Substitution Generation (SG) and Substitution Selection (SS). Each task is implemented with two different methods. For SG, Word2Vec, augmented with Part of Speech (POS) tagging, is compared against a deep learning-based approach–Masking Language Model (MLM). In SS, cosine similarity is computed between the candidates, generated from the Word2Vec approach, and the content words in the sentence, while MLM-generated candidates are ranked using a decoder-only language model through prompt-based selection. The MLSP 2024 Dataset (Shardlow et al., 2024) is used to test the generated and selected candidates. The metrics in this study are Potential@k, Recall@k, and Precision@k. The results indicates that the candidates generated using MLM achieved higher score in the three metrics than the candidates of Word2Vec. The findings highlight the promising potential of language models in tackling the challenges of LS pipeline.

1 Introduction

Reading comprehensively can be hindered when complex and unfamiliar words are present, thereby leading to obstacles in reading. It is affirmed that people with a small linguistic repertoire and people with special conditions like Dyslexia and Aphasia (Paetzold and Specia, 2017) are susceptible to problems in comprehension. Therefore, the task of lexical simplification aims to make text more accessible by identifying difficult words and replacing them with simpler synonyms or phrases without changing the pragmatic meaning of the sentence. For example, to reduce complexity, a word such as "preposterous" can be changed to "unreal" for simplicity. A LS system typically operates in a four-task pipeline. The first task identifies the complex word in a given sentence, while the next task generate candidates for the complex word selected. Then, the third task selects the top k of the generated candidates. The task of ranking the selected candidates surfaces as the last task. In the survey of (North et al., 2025), the authors argue that with modern deep learning apporaches, Substitution Selection and Substitution Ranking are combined as a one task. Nonetheless, by convention, they are four tasks. In this paper, the focus is placed on Substitution Generation and Substitution Selection, with two methods for each task. For generating, Word2Vec is used alongside with Masking Languge Model (MLM). Semantic similarity and prompting a language models are two distinct methods used or selecting the generated candidates. Although deep learning approaches achieved higher results than Word2vec, they yielded in some ill-formed synonyms and out of context candidates.

2 Related Work

The field of Lexical Simplification, and particularly its Substitution Generation (SG) and Substitution Selection (SS) sub-tasks, has long attracted interest from researchers. For example, (Young, 1999) investigated simplifying reading materials for Second Language (SL) learners, The results showed that simplification did not, in most of the cases, achieve high scores, and, surprisingly, authentic texts yielded better outcomes. Nonetheless, as time went by, new LS methods came into the scene. For example, early approaches for generating substitutes, relied heavily on WordNet database (Soergel, 1998) to extract synonyms of complex words. (Paetzold and Specia, 2017) argued that many researchers resorted to WordNet for LS tasks. However, WordNet is not as an optimal approach since it does not count for all the complex words in the English language, not generating reasonable candidates in most cases (Shardlow, 2014). A shift towards semantics-based approaches came with the use of word embeddings, (Paetzold and Specia,

2016) proposed a pipeline dedicated for non-native English speakers. They used context-aware word embeddings, aided with POS tags, for generating substitutes, and trained it on a large corpora of movie subtitles. Their system outperformed prior state-of-the-art LS models. Another study that used word embeddings is (Glavaš and Štajner, 2015). They used the GloVe vectors (Pennington et al., 2014). Upon the advent of the transformers architecture, recent surveys have focused on deep learning approaches for SG. For example, (North et al., 2024) highlights a key approach for substitution generation–Masked Language Modeling (MLM). MLM is about masking the complex word with a special token [MASK] and asking the language model to predict the most suitable word, based on the context. (Qiang et al., 2020) introduced LS-Bert, which uses BERT's masked language model to generate context-sensitive substitutes. When evaluated on datasets like LexMTurk (Horn et al., 2014), BenchLS, and NNSeval (Paetzold and Specia, 2016b), LSBert, in the TSAR-22 sharedtask (Saggion et al., 2022), achieved F1-scores of 0.259, 0.272, and 0.218, respectively. This outperforms the static-embeddings approaches which lacks understanding of contextual nuances. For Substitution Selection (SS), (Paetzold and Specia, 2015) used word embeddings to compute the semantic similarity between the candidates and the content words in the sentence, retrieving only the words with the highest score. Another vector-based approach is what (Biran et al., 2011) proposed of comparing co-occurrence vectors of the target word in context with those of each candidate using cosine similarity. With transformers architecture, prompt learning has entered the LS landscape. That is, (North et al., 2023), after filtering steps, fed the candidates to ChatGPT 3.5. to select, with the use of adjectives, the simplest, best, or most similar to identify the most appropriate replacement. This approach is deemed to be the most promising one amidst the previous selection methods. However, there is still a venue for further investigation despite these advances. Therefore, in this study, a Word2Vec-based pipeline is compared with a deep learning-based pipeline for both SG and SS. For SS within the deep learning pipeline, rather than using GPT-based models, this work explores Mistral-7B-Instruct-v0.1 (Jiang et al., 2023). Also rather than using MLM on the aforementioned datasets, the study examines the MLSP dataset (Shardlow et al., 2024).

3 Methodology

This section delves into the steps of conducting this experiment, starting with a description of the dataset used to evaluate the generating and selecting models. Then, pre-porcessing of the dataset is investigated. After that, an extensive explanation is given to deconstruct the models implemented to tackle the task of LS.

3.1 Explanatory Data Analysis

The dataset used here is the English Combined Lexical Simplification (LS) test set from the MLSP 2024 Shared Task (Shardlow et al., 2024). The dataset comprises 570 instances, each with a target complex word, level of complexity, and multiple human-annotated gold substitutions. An example of how the dataset is structured in the appendix A.1. Across the dataset, 47% of all unique gold-substitutes appear more than once in the dataset. (see Appendix Figure 1 for the most repeated substitute words). Another aspect noticed in the dataset is that the majority of complex words fall between 0.10 and 0.31, indicating most words are relatively low in the complexity scale, as shown in Figure 2 in the Appendix. Although the dataset possesses 30 columns for substitutions, most complex words have between 8 and 10 substitutions, with few cases exceeding 10, as shown in Figure 3. Given that, the dataset provides a reasonable and consistent range of candidate substitutions, making it suitable for evaluating both substitution generation and selection models.

3.2 Data Preprocessing

The preprocessing was kept simple and focused. The dataset was filtered to keep only the important parts: the full sentence (`context`), the complex word that needs to be simplified (`target`), and the list of human-provided substitutions (`substitution_1` to `substitution_30`). Tokenization was implemented as needed for the Word2Vec language model.

3.3 Substitution Generation

Two distinct methods are used to generate substitute candidates–one is based on static and pragmatics-deprived embeddings and the other is based on contextual embeddings. This section analyses these two approaches.

3.3.1 Word2Vec Embedding and POS Filtering

a pre-trained Word2Vec model, fed with Google new corpus, is leveraged to potentially yield synonyms. The notion is that, given a complex word, the model generate the top-n nearest neighbors. The number of generated candidates is set to 10 to have various words. This raw list, however, has unfiltered POS tags, potentially yielding ill-formed candidates, such as verbal candidates for adjectival target word. Therefore, a POS filtering is applied. That is, each target word is tagged with its POS using the `NLTK's POS tagger`. This in return filters the candidate list to only include words with the same POS tag. An example of the substitute candidates of the target word `equals` is a list of words that are similar in meaning such as ['equates', 'translates', 'means', 'begets', 'implies', etc.]. To provide an intuitive understanding of the semantic clustering of Word2Vec-generated substitutes, a plot is included in Figure 4. The Word2Vec method provides an intuitively straightforward approach to generate words. However, static embeddings do not count for contextual understanding beyond the target word itself.

3.3.2 Masked Language Model

This generative method is based on the `bert-base-uncased` model (Devlin et al., 2019). The pipeline is imported from Huggingface with the task `fill-mask`. the function replaces the complex word with a [MASK] token, and then, based on the context, the model predicts the top n substitutes that are pragmatically and syntactically similar to the target word. the top N predictions (again N = 10) are retrieved from the softmax output of BERT. There is no need for any pre-processing steps when such a Transformers-based model is used. The model, with bidirectional attention mechanism, understands how to match the POS of the complex word with its substitute candidates. The candidates are context-sensitive, outperforming the candidates of Word2Vec. Some examples of the MLM approach are shown in table 1

Both models were instructed to generate 10 candidates per one target word. However, while the MLM model genereted 10 words for each word, the Word2Vec model often fell short, sometimes producing far fewer—or even zero—candidates, as shown in Figure 5. This drawback in the Word2Vec model is contributed to the strict part-of-speech (POS) filtering and the rarity of certain target words in Word2Vec.

3.4 Substitution Selection

Given two distinct corpora of generated substitutes, two different approaches are used to selected the best substitutes from the generated candidates.

3.4.1 Word2Vec-based candidates

the approach of (Paetzold and Specia, 2015) for selecting substitutes is employed in this study. The overall notion of the approach is about computing the semantic similarity between the context words (sentence) of the target word and each Word2Vec-based generated substitutes. This starts with vectorising the content words of the sentence–without the target word–and averaging the embeddings of all content words (i.e., words that are not stop-words or punctuation). This context vector serves as a semantic summary of the sentence. Then, each candidate generated from the Word2Vec-based Substitution Generation step is then embedded using the same model, and its cosine similarity to the context vector is computed. words that semantically aligns the most with the summary vector of the content words are given the highest rank. words are then ranked in a descending order. To select the most similar candidates after the ranking, a threshold of 50% is placed to nominate the last set of candidates.

3.4.2 BERT-based candidates

BERT already provides candidates in order of likelihood, which often correlates with how well they fit the context. However, in some cases, the model hallucinates, generating antonyms, prepositions, and unrelated words. Therefore, prompting the open-source 7B parameter LLM (Mistral 7B) is leveraged to select the best candidates generated from the MLM approach. The idea is to prompt this model with a structured instruction that lists the context sentence, the target word, and the candidate list. Mistral was prompted with the following: "<s>[INST] Given the sentence: "context" What is the best replacement for the complex word target" in this list? {candidate_str} Respond with one word only. Do not explain. [/INST]"

The reason the prompt is explicitly imperative is because Mistral would either define the candidate words, giving it a simple prompt like "what is the best replacement", or hallucinate by returning numbers. Moreover, the hyperparameters of the model

are fine tune to generate the most plausible results. (See Table 2 for the hyperparameters)
After crafting the prompt, the model selects the best candidate for all the instances in the dataset. requiring a long time, the dateset is split into three batches to save time and GPU energy.

4 Results and Evaluation

To quantitatively evaluate the performance of the generating and selecting pipelines, three LS-based metrics are used, $Potential@k$, $Precision@k$, and $Recall@k$. Potential investigates if there is at least one correct gold substitute in the $top - k$ generated and selected words. While precision calculates the average proportion of how many k words are valid. But recall calculates how many of the gold substitutes are found in SS and SG.

4.1 Evaluation of the SG pipeline

Table 3 summarizes the performance of the two SG methods on the MLSP 2024 evaluation dataset. The Masked Language Model (MLM)-based generator achieves higher values in all metrics than Word2Vec. The MLM approach attains a $Potential@10$ of 0.577, opposed to 44.2% for Word2Vec. It also yields higher $Recall@10$ (about 0.201 vs 0.115). The Precision@10 is likewise higher for MLM (0.098 vs 0.072). MLM achieving higher scores confirm that the context-sensitive method can generate more relevant and comprehensive candidate lists than the static embedding approach. The Word2Vec generator, even when aided by POS filtering, often misses suitable synonyms or proposes related words that are not true synonyms, which lowers both its recall and precision. In contrast, the MLM approach leverages the contextual nuances, fed by the attention mechanism, allowing it to suggest more appropriate synonyms. For example, in the sentence, "Buddhism and Jainism share many common themes and practices, including an emphasis on nonviolence and a belief that one achieves liberation...", the MLM generator returned a correct context-based candidate for the word "liberation" such as "salvation", focusing on simplicity. while the Word2Vec counted for neighboring similar words but not simpler such as "emancipation.

4.2 Evaluation of the SS pipeline

Table 4 summarizes the performance of the two SS methods on the MLSP 2024 evaluation dataset. the two SS pipelines do not share the same threshold in the metrics because the semantic-similarity SS approach only ranks words as the most similar without actually selecting candidates. The prompt selector select only one word among the generated list. That being said, a threshold of 50% is placed for the word2vec approach and Mistral returns only one word. The results show that the Word2Vec ranker achieved higher results across the three metrics compared to the prompt selector. To elaborate, the ranker, aided with the threshold, selects more than one candidates in most of instances, boosted to have a higher score although the ranker would erroneously rank unrelated words or antonyms as the most similar in some cases. However, even with the low score for the prompt selecting method, Mistral achieved to select correct substitutes. For example, in the sentence "Everything that you study is geared to prepare you for organic syntheses and other chemical transformations performed in the lab.", Mistral accurately selected conducted from the MLM-generated list to be replaced with the target word "performed".
In short, not only do the selected candidates rely on the selecting methods but also requisitely on the generating methods. If the generating methods are abstract and simple, the overall performance will be hurt, selecting poor candidates.

5 Error Analysis

Although the quantitative analysis in the previous section deconstruct major-structure results, it is still necessary to go over specific ill-formed results that happened in both generation and selection.

5.1 Errors in Word2Vec-based SG

the first failure encountered in this model is that 9.82% of the complex words in the dataset had no substitutes generated. This is contributed to many factors. One of which is that the POS filtering added more strictness to only match the target word's form. Also, the target word itself is rare, as shown in table 5. Words like "nucleophiles", "electronegative", and "parabolic" are difficult to be found in the vocabulary of Word2Vec. Another drawback detected is that in many cases the generated candidate is antonymous to the target word. It is detected to be 3.51% of the Word2vec-generated candidates are antonyms as shown in Table 6.

5.2 Errors in MLM-based SG

Upon inspection of the MLM-generated substitutes, several types of errors were identified. In some cases, the generated output included subtokens. For instance, the tokens `##oon`, `##ers`, and `##ing` appeared as a substitute where no complete word or meaningful context aligned with it. These cases often arise due to the model's reliance on subword tokenization (e.g., WordPiece), which can produce partial tokens not semantically valid in isolation. Another flaw of the MLM model is that model is heavily thematic deemed. To elaborate with an example: *"After the war, Hitler remained in the army and after receiving intelligence and oratory training, became an intelligence official tasked with infiltrating political parties and reporting to his superiors on their activities."*
Here, the generated substitute for *oratory* included terms like *['military', 'combat', 'practical', 'police', 'political', 'officer', 'diplomatic', 'weapons', 'operational', 'further']*, which reflect thematic bleeding from other parts of the sentence such as *war*, *army*, and *intelligence*.

5.3 SS Errors in general

Since the two selecting pipelines rely on the generated candidates, the Substitution Selector (SS) failed not only because of poor ranking, but because the generator produced no valid substitutes. One example is that the target words, not substituted with any candidate in the Word2Vec-based generator, ended up receiving no substitute at all. That tells that the Word2Vec model failed in generating candidates for rare words as mentioned in table 5. Also, the Word2Vec ranker erroneously ranked unrelated words as the most similar. For example, after the Word2Vec generated correct substitutes for "alleviate", the ranker misranked the word reduced as the most similar to the content words. See table 7 for more examples. This is justified by the fact that w2v leans toward the most neighboring words not the simplest. The Mistral model also hallucinated when selecting. For example, after the MLM generated candidates, the Mistral model refused to select candidates as shown in table 8. Also, Mistral in many cases would hallucinate by choosing words that are not in the MLM-based list, as proven in table 9. One reason could be contributed to this type of error is the `temperature` hyperparameter, which was 0.7. That increases the chance for the model to be creative, possibly compelling it to choose words that are context-driven even if they are not in the list from which the model should have chosen.

6 Ethical Reflection

Lexical Simplification offers societal benefits, simplifying complex words for people with a small linguistic repertoire or people with dyslexia and aphasia. However, an ethical concern is the risk of *oversimplification*. By replacing complex words with simpler ones, there is potential that the replaced words might loose nuances and critical semantic information. Another risk involves the misuse of simplification tools. If applied recklessly to critical documents, such as the ones of the medical and legal domains, oversimplification might omit vital information. Moreover, embeddings-based models introduce concerns about bias and accuracy. These systems may produce substitutes that are biased or contextually poor if the data used for training is driven with imbalance and misinformation. Awareness of these limitations and efforts to mitigate bias are important to use these models in a safe environment, not carrying repercussions.

7 Conclusion

This study compared static and contextual embeddings across two stages of the lexical simplification pipeline–Substitution Generation (SG) and Substitution Selection (SS). Results showed that the MLM generator consistently outperformed Word2Vec in *Potential@10*, *Recall@10*, and *Precision@10*, highlighting the importance of contextual nuances when generating candidate substitutes. In the selection stage, the Word2Vec-based semantic–similarity ranker achieved higher scores than the prompt-based Mistral selector, largely because it could allow multiple candidates per instance. Nonetheless, error analysis showed that both selectors are ultimately constrained by the quality of their generated candidate lists. Word2Vec missed many valid substitutes for rare or morphologically complex words, while Mistral occasionally hallucinated or failed to return an answer.
Finally, further research should investigate fine tuning the hyperparameters of the Mistral model to mitigate hallucinations. Also, domain-specific encoder-only language models should be used for generating candidates.

References

Shardlow, M., Alva-Manchego, F., Batista-Navarro, R., Bott, S., Calderon Ramirez, S., Cardon, R., François, T., Hayakawa, A., Horbach, A., Huelsing, A., *et al.* (2024). An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework. *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI).*

Paetzold, G.H. and Specia, L. (2017). A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60, 549–593.

North, K., Ranasinghe, T., Shardlow, M. and Zampieri, M. (2025). Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, 63, pp.111–134. Available at: https://doi.org/10.1007/s10844-024-00882-9.

Young, D.J., 1999. Linguistic simplification of SL reading material: Effective instructional practice? The Modern Language Journal, 83(3), pp.350–366. Available at: https://www.jstor.org/stable/330258 [Accessed 19 Apr. 2025].

Specia, L., Jauhar, S.K. and Mihalcea, R. (2012). SemEval-2012 Task 1: English lexical simplification. *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pp.347–355.

Shardlow, M., 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1), pp.58–70.

Glavaš, G. and Štajner, S., 2015. Simplifying lexical simplification: Do we need simplified corpora? *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp.63–68.

Pennington, J., Socher, R. and Manning, C.D., 2014. GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.1532–1543.

North, K., Ranasinghe, T., Shardlow, M. and Zampieri, M., 2024. Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, 63, pp.111–134. Available at: https://doi.org/10.1007/s10844-024-00882-9 [Accessed 19 Apr. 2025].

Horn, C., Manduca, C. and Kauchak, D., 2014. Learning a lexical simplifier using Wikipedia. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, USA, 22–27 June 2014, pp.458–463. Paetzold, G.H. and Specia, L.,

2016. Benchmarking lexical simplification systems. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 23–28 May 2016, pp.3074–3080. Paetzold, G.H. and

Specia, L., 2016. Unsupervised lexical simplification for non-native speakers. In: *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, 12–17 February 2016, pp.3761–3767. Aumiller, D. and Gertz, M.,

2022. Investigating lexical simplification with GPT-3: Promising results using zero-shot and few-shot learning. In: Proceedings of the 1st Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022). Abu Dhabi, United Arab Emirates, 12 December 2022. Stroudsburg, PA: Association for Computational Linguistics, pp.247–259. Available at: <https://aclanthology.org/2022.tsar-1.28/> [Accessed 19 Apr. 2025].

Paetzold, G.H. and Specia, L., 2015. Lexenstein: A framework for lexical simplification. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, Beijing, China, 26–31 July 2015. Stroudsburg, PA: Association for Computational Linguistics, pp.85–90.

Biran, O., Brody, S. and Elhadad, N., 2011. Putting it simply: A context-aware approach to lexical simplification. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, Portland, Oregon, 19–24 June 2011. Stroudsburg, PA: Association for Computational Linguistics, pp.496–501.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Renard Lavaud, L., Saulnier, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T. and El Sayed, W., 2023. Mistral 7B. [online] Available at: https://arxiv.org/abs/2310.06825 [Accessed 20 Apr. 2025]

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. [online] Available at: https://arxiv.org/abs/1810.04805 [Accessed 21 Apr. 2025].

A Appendix

Context	Target	MLM substitutes (top–10)
All other things being equal, nucleophiles are generally compared to each other in terms of *relative* reactivity.	relative	their, chemical, the, relative, nuclear, molecular, net, overall, total, biological
As adjectives, their endings will vary according to the nouns they modify.	according	according, relative, based, referring, related, in, following, as, compared, accordingly
By the 1960s more mail was being franked by postage meter than with traditional *adhesive* stamps.	adhesive	postage, british, postal, paper, revenue, mail, german, chinese, european, commercial

Table 1: Three examples with their targets and full MLM substitute lists.

Hyperparameter	Value	Justification
`max_new_tokens`	5	Limits output 1 word.
`temperature`	0.7	to boost creativity in selection.
`top_k`	20	Samples from top 20 likely tokens to avoid low-probability noise.
`top_p`	0.9	Enables nucleus sampling for diversity without sacrificing quality.
`do_sample`	True	Allows probabilistic sampling instead of deterministic decoding.
`repetition_penalty`	1.2	Prevents repetitive outputs.

Table 2: Decoder LLM generation hyperparameters used for substitution selection.

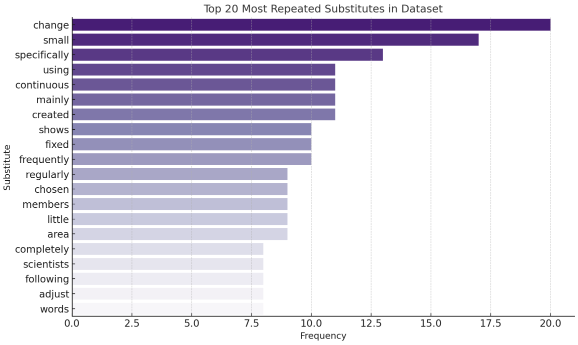

Figure 1: Top 20 most repeated substitutes in the MLSP 2024 English test set. This overlap shows the frequent reuse of simplified vocabulary across instances.

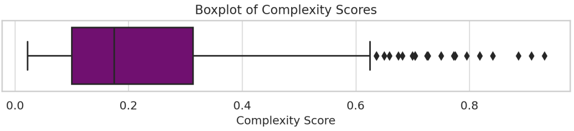

Figure 2: Boxplot showing the distribution of complexity scores in the MLSP 2024 English test set. The purple box highlights the interquartile range (Q1–Q3), where the majority of scores fall between 0.10 and 0.31.

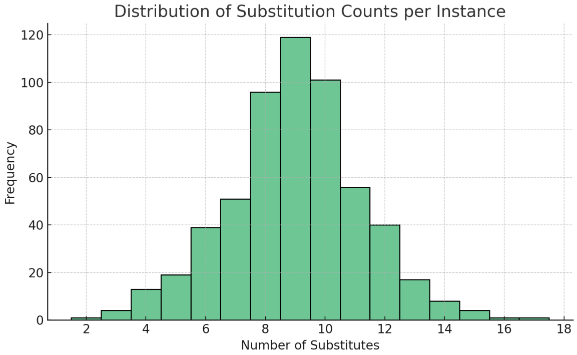

Figure 3: Histogram showing the distribution of the number of gold substitutes per complex word in the MLSP 2024 test set. The majority of instances possesses 8 to 10 substitute candidates.

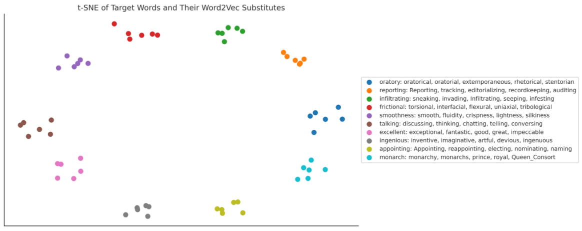

Figure 4: t-SNE visualization of 10 target words (colored clusters) and their top 5 Word2Vec-generated substitutes. Each color corresponds to a different target word, and the legend shows the word along with its generated candidates.

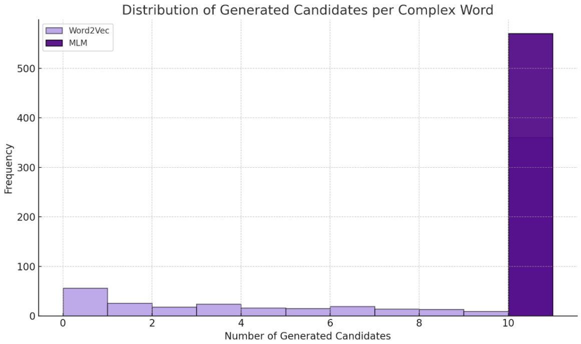

Figure 5: Distribution of the number of generated candidates per complex word for Word2Vec and MLM. While MLM consistently produces 10 candidates per target, Word2Vec frequently returns fewer due to strict filtering.

SG Method	**Potential@10**	**Precision@10**	**Recall@10**
W2V	0.442	0.072	0.115
MLM	**0.577**	**0.098**	**0.201**

Table 3: Comparison of substitution generation (SG) methods on the evaluation dataset. The MLM-based method outperforms the static embedding approach in all metrics (higher is better).

SS Method	**Potential@p**	**Precision@p**	**R@p**
W2V Context-Ranking (Top 50%)	0.334	0.113	0.095
Prompt-based Selection (Top 1)	0.208	0.021	0.045

Table 4: Comparison of substitution selection (SS) methods. The Word2Vec selector uses contextual similarity and retrieves a top proportion of substitutes. The prompt-based selector selects from a list of 10 generated candidates using a large language model.

Table 5: Examples of Word2Vec failing to generate substitutes

Context	**Target**
All other things being equal, nucleophiles are more reactive when the atom bearing the charge is less electronegative.	nucleophiles
Although the Jews were the favored targets and principal victims, the Nazis and their collaborators also murdered millions of others.	hear
Among these technologies was nuclear fission, the discovery that led to the development of nuclear power and nuclear weapons.	electronegative
Another separate book covers Cascading Style Sheets (CSS) in detail.	addresses
Another separate book covers Cascading Style Sheets (CSS) in detail.	handle
As adjectives, their endings will vary according to gender, number, and case.	vary
As the single parabolic reflector achieves a greater gain, it also narrows the beamwidth.	parabolic
Attachment to the fibre is attributed, at least in part, to ionic interactions.	fibre
British trade officials took it as a first priority to populate the colony with British settlers.	populate
Click the attached MAC or WIN button to download the latest version of the software.	download

Target	**Generated Antonym**	**Context**
father	mother	As Governor, he followed such a strict policy of appointing only fellow Republicans to office...
queen	king	As well as being queen of the United Kingdom of Great Britain and Ireland, she was also the first mo...
empirical	theoretical	Configuration can also be assigned on the purely empirical basis of the optical activity.
willing	unwilling	If possible locate a beekeeper in your area that may be willing to guide you along or even mentor you...
unfavorable	favorable	In many annual species, only the seed exists during unfavorable dry or cold conditions.
unfavorable	Favorable	In many annual species, only the seed exists during unfavorable dry or cold conditions.
unfavorable	Very_Favorable	In many annual species, only the seed exists during unfavorable dry or cold conditions.
northern	southern	In northern Europe (north Germany, Netherlands, and France), the middle class tended to be Protestant...
sufficiently	insufficiently	In order to obtain this field, the room has to be sufficiently reverberant and the frequencies have to...
frequently	infrequently	In organic chemistry, carbon is very frequently used, so chemists know that there is a carbon atom at...

Table 6: Examples of antonym substitutions generated by Word2Vec

Context (truncated)	Target	Mis-ranked gold sub(s)	w2v top pick(s)
After Ron nearly dies drinking poisoned mead that was apparently intended for Pr…	distraught	distressed	angry
After the war, Hitler remained in the army and after receiving intelligence and …	reporting	documenting	auditing, disclosing
After the war, Hitler remained in the army and after receiving intelligence and …	infiltrating	intruding, penetrating	assimilating, invading, rapidly_proliferating
All other things being equal, nucleophiles are generally compare…	compared	contrasted	Compared, approximated, decreased
Also, the frictional coefficient varies greatly depending on wha…	varies	alters, differs, ranges	Pricing_varies, determines, fluctuates
Among these technologies was nuclear fission, the splitting of a…	breakthroughs	innovations	Technological_breakthroughs, nanotools, tech_nologies
An ingenious alphabet allowed the Maya to record information on …	excellent	exceptional, outstanding	good, great, marvelous
An ingenious alphabet allowed the Maya to record information on …	ingenious	innovative, inventive	alchemic, artful, imaginative
As Governor, he followed such a strict policy of appointing only…	selecting	naming, nominating	electing, reappointing, renominating
As well as being queen of the United Kingdom of Great Britain an…	monarch	royal, ruler	Queen_Consort, king, queen

Table 7: Examples where word2vec mis-ranked gold substitutes

Context (truncated)	Target	MLM generated substitutes (truncated)	Mistral output
It repeatedly shows illustrations of Scouts in sunny mountain sc…	whatsoever	[’on’, ’attached’, ’underneath’, ’worn’, ’removed’, ’around’…	None
The detriments outweigh the benefits and there are no current st…	vouchers	[’they’, ’these’, ’both’, ’none’, ’those’, ’drugs’, ’treatme’…	None
When a text fails to support this rule, it requires even a dilig…	fails	[’fails’, ’attempts’, ’appears’, ’tries’, ’exists’, ’seems’,…	Refuses

Table 8: Cases where the Mistral selector failed to return a substitute

Context	Target	MLM generated substitutes	Mistral pick
Drill them out using a very slow drill speed, or use a screw ex…	extracting	pull, remove, loosen, remove_it, remove_them, take_out, take, pull_it, pull_them, unscrew	Stuck
Even under ideal conditions there can still be discrepancies bet…	focusing	focusing, thermal, optical, visual, image, illumination, angular, horizontal, viewing, temperature	Optimizing
The resulting structure leaves the surface porous and therefore…	smoothness	pressure, pressures, density, humidity, friction, strain, ph, motion, concentration, concentrations	Surface roughness
The whole instrument is tuned so that it can be played accordin…	harmonic	pitch, rhythm, beat, note, key, interval, tone, melody, frequency, chord	Melody
The player has to calculate the force and direction of each sho…	aiming	aiming, shooting, angle, shot, trajectory, power, force, strength, speed, velocity	Targeting
Because his mating apparatus is barbed, like the stinger of a w…	stinger	antennae, body, penis, head, apparatus, abdomen, tail, collar, legs, one	Stinger
Although the amount of available storage is limited, careful ma…	storage	memory, space, disk, capacity, ram, flash, hard_drive, database, server, cache	Storing
The legislator argued that any measure which restricts freedom …	regulation	regulation, law, rule, policy, legislation, statute, code, directive, act, ordinance	Regulate
During the medieval period, monasteries often served as centers…	scholarship	learning, study, teaching, education, research, knowledge, reading, writing, literacy, culture	Academic
Hard lines are often made to be pressed lightly at first, so th…	sketching	drawing, outlining, tracing, marking, drafting, penciling, shading, illustrating, rendering, doodling	Sketch

Table 9: Sample cases where the Mistral selector’s chosen substitute is absent from the MLM candidate list

A.1 Appendix for the Examples

Context: “After Ron nearly dies drinking poisoned mead that was apparently intended for Professor Dumbledore, Hermione becomes so distraught that they end their feud for good.”

Target: *distraught*

Gold substitutes: worried, troubled, distressed, upset, shaken