

A Hybrid Transformer-Based System for Multilingual Translation with Domain-Specific Terminology Explanation

Oyebamiji Micheal Tomiwa

Computer Science, Faculty of Computing, University of Ibadan, Ibadan, Oyo, Nigeria

E-mail: oyebamijimicheal10@gmail.com

(Received: 31 May 2026, revised: 28 Jun 2026, accepted: 29 Jun 2026)

Abstract

Communication across languages becomes particularly challenging when specialized terminology is involved because literal translations often fail to convey the precise meanings essential in professional contexts. Current translation systems have problems with translating technical terms from specific domains like law, medicine, or finance, such that they generate translations that lack explanations of what technical terms actually mean. This creates significant problems in professional settings where misunderstanding specialized vocabulary can have serious consequences. This research developed a hybrid transformer-based system that addresses these limitations by translating specialized terminology across legal, medical, and financial domains while providing contextual explanations in the target language. The system used two transformer models: Sentence-BERT to detect special terms and MarianMT to translate. Data collection involved extracting 7,100 specialized terms with definitions from authoritative sources, alongside 70,000 parallel sentence pairs for each of three language pairs: English-Spanish, English-French, and English-German. The CRISP-DM framework was used for development from problem definition to deployment. The English-Spanish model achieved 59.26 BLEU, English-French achieved 38.76 BLEU, and English-German achieved 25.68 BLEU. Term extraction achieved 92.8-96.5% average accuracy across domains, with perfect performance on exact matches and 79-94% accuracy on misspelled or incomplete terms.

Keywords: Machine Translation, Transformer Models, MarianMT, Sentence-Bert, Domain-Specific Terminology, Multilingual NLP.

1. INTRODUCTION

The basis of human communication is language [1]. It is a medium that is used to communicate ideas, connect with people, and share emotions. According to the Oxford Research Encyclopedia of Linguistics, there are about 7,000 languages spoken around the world today [2]. This language diversity is testimony to the richness of human cultures, but also poses great challenges to intercultural dialogue. In this era of globalisation, the world is more connected than ever, and accurate and quick translations are more crucial than ever to allow people from different backgrounds to interact effectively.

It was in the 1950s that the process of translation started, with the first systems using a handwritten dictionary and language norms, such as Transfer-Based or Interlingual approaches and Direct Machine Translation, which emphasized word-for-word translation [3]. In the 1990s, the field shifted to the statistical era, based on probabilistic models learned from large parallel corpora. The Statistical Machine Translation (SMT) systems used statistical methods to learn

the trends of translation from data [4] and achieved significantly better results than rule-based systems, particularly for language pairs where there was a large amount of training data. Using deep learning techniques, the current neural and transformer era, which began in 2014, has reshaped machine translation. With sequence-to-sequence models and attention mechanisms, complex linguistic patterns could be learned end-to-end [5]. The Transformer architecture [6] introduced by Vaswani et al. has been shown to be state-of-the-art by replacing the recurrent architecture with self-attention. More recently, translation models like MarianMT [7] and T5 [8] have further advanced translation quality for a variety of language pairs.

Despite these developments, there are still domain-specific translation problems that remain prominent. General-purpose translation systems are not as effective when dealing with specialised fields like law, medicine, and finance. Medical texts contain complex medical and pharmaceutical terms which must be translated accurately for patient safety; legal documents include terms with precise meanings, often with no direct equivalents in other languages; and financial documents



often employ specialised terms where even a minor mistranslation can carry serious economic consequences [9].

One of the major gaps that is not sufficiently addressed is the lack of contextual explanations in translation outputs. As illustrated in Figure 1 and Figure 2, even well-established general-purpose translation platforms struggle with legal expressions like "challenge to the array." Google Translate renders the phrase as "objection to the list of evidence" while DeepL translates it as "challenge to the jury," as shown in Figure 2. Neither platform conveys that the expression refers specifically to a legal objection to the entire jury panel selection process, typically raised on procedural grounds. Both translations would leave a legal professional or their client with a fundamentally misleading understanding of what was filed. The users receive no information about the meaning of the term in its domain-specific context, even when a term is translated in a way that sounds plausible, which is particularly problematic in professional contexts where misreading terminology can have serious consequences [10].

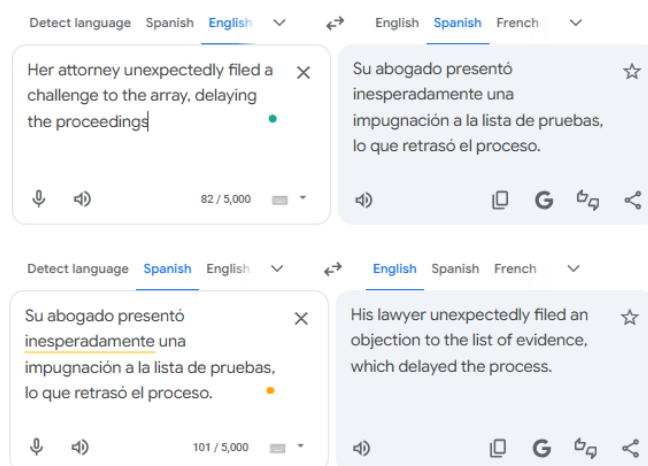


Figure 1. Google Translate Output for the Legal Term "Challenge to the Array," Which Gives a Surface-Level Translation That Does Not Reflect the Term's Legal Meaning

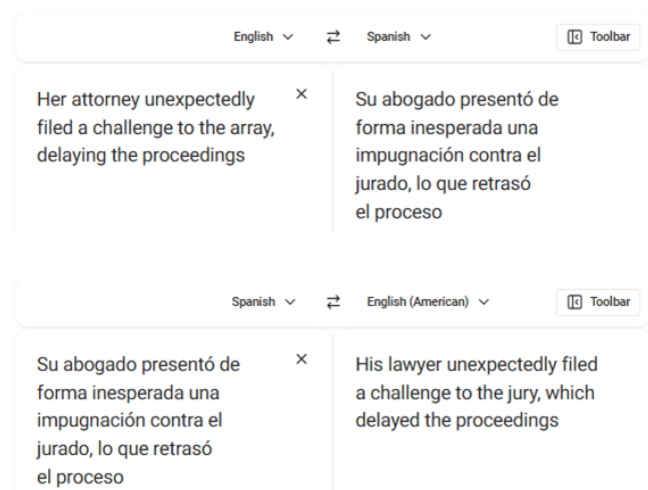


Figure 2. DeepL Output for the Same Legal Expression, Producing an Alternative But Equally Incomplete Translation Without Domain-Specific Context

In this paper, we present a hybrid system that combines MarianMT for translation and Sentence-BERT (SBERT) [11] for domain-specific term identification and explanation retrieval. The system focuses on three domains: legal, medical, and financial, and three language pairs: English-Spanish, English-French, and English-German. The main contribution is a text translation pipeline that translates texts from one language to another, but also detects technical terms and explains them in the target language, a gap that is not sufficiently covered in the existing literature.

II. EXISTING WORK

There have been several studies that have tried to tackle domain-specific translation problems, with some successes and drawbacks. Müller et al. [12] investigated domain robustness in Neural Machine Translation (NMT) systems and reported average out-of-domain BLEU improvements of 1.5 for German-English and 2.7 for German-Romansh, but they did not consider the issue of semantic adequacy for specialised terminology. In-context learning for domain-specific translation was investigated by Agrawal et al. [13], who reported 59.57 BLEU for the medical domain and 46.64 BLEU for the legal domain, though their method was sensitive to noisy examples and they only evaluated their results using BLEU metrics. Khandelwal et al. [14] proposed k-nearest neighbour machine translation, which achieved an average domain adaptation gain of 9.2 BLEU points without additional training, though the method required substantial storage and offered limited interpretability for retrieved results.

In the financial field, Alghamdi et al. [15] fine-tuned models such as OPUS, NLLB, and ChatGPT-3.5 on financial texts and demonstrated that the domain adaptation method worked to enhance translation accuracy, but was domain-specific and needed a large volume of in-domain data. In the medical terminology field, Limsopatham and Collier [16] used phrase-based machine translation and word vector representation to normalize informal medical terms from social media, with an MRR-5 score of 0.2638, but their work was limited to the adverse drug reaction domain.

In addition to these retrieval and adaptation techniques, Zeng et al. [17] explored a multi-domain NMT system that incorporated domain-specific token weights during training, and they achieved a BLEU score of 45.82 to 55.03 in the legal domain; however, they did not attempt to explain the meaning of specialised terms in a semantic way. Poudel et al. [18] focused on legal translation between English and Nepali, and trained a Transformer-based system on 125,000 parallel legal texts, which scored 6.63 and 7.98 on BLEU, respectively, but observed that even domain-trained models performed poorly with complex legal jargon without jurisdiction-specific adaptation.

In all of these studies, a consistent gap is evident. Much work has been done on improving the accuracy of the translation in a specific domain or a specific language pair, but the general problem of giving explanations for special

terms in a context is not yet sufficiently studied. Most systems make the assumption that a translated word is all that is needed and neglect the requirement for users to know the meaning of a technical term in context. This limitation motivates the proposed hybrid approach to MarianMT, which is based on translation and domain-specific term explanation in legal, medical, and financial domains.

III. RESEARCH METHODOLOGY

A. Framework and Design

The development process was guided by the CRISP-DM framework, which is originally a data mining framework but has been adapted in this project for the NLP development process. The development was done across six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. During the business understanding phase, the major problem was discovered: the existing translation systems are not able to cope with the translation of domain-specific terminology sufficiently, and they don't offer any explanation for domain-specific terms. The data understanding phase involved looking at the sources of legal, medical, and financial terminology. The data preparation stage included cleaning, tokenisation, and structuring of all corpora into usable formats for model training.

B. Data Collection

Three domain-specific corpora were created. The legal corpus was taken from the Handbook of Legal Terminology published by the University of Mississippi [19], which contains 750 terms with definitions. The medical terminology was extracted from Merriam-Webster's Medical Dictionary, Third Edition [20] which provides 5,100 medical terms with clinically accurate definitions. Two sources were used to provide financial terminology: the World Bank Glossary of Finance and Debt [21] and the Financial Glossary of City College of San Francisco [22], which provides 1,250 financial terms and their definitions. For each corpus, pipelines were created to parse PDFs using PyMuPDF.

For translation model training, parallel corpora were assembled for three language pairs. For English-Spanish, a Kaggle dataset of 139,013 sentence pairs was used, from which the last 70,000 entries were selected. As shown in Figure 3, source English sentences average 7.8 words while target Spanish sentences average 7.5 words, with most sentences concentrated between 5 and 10 words. For English-German, 70,000 examples were extracted from the WMT14 dataset via Hugging Face. Figure 4 shows that source English sentences average 25.3 words while target German sentences average 23.3 words, a substantially broader distribution that has direct implications for how the models are trained and how their BLEU scores should be interpreted. Each language pair was split into training (56,000 pairs), validation, and test sets.

C. Model Development

The system was based on two transformer models. To extract specialised terms, Sentence-BERT (all-MiniLM-L6-

v2) was used to generate semantic embeddings. Two types of embeddings were calculated for each domain term: enriched embeddings that combine the term and its definition to provide richer context, and term-only embeddings to provide exact matching. When a user provides input text, the system breaks it down into combinations of phrases of varying length, creates embeddings, and calculates a hybrid similarity score based on semantic similarity (40%), direct term matching (35%), word overlap (15%), and length similarity (10%). These weights were assigned based on the relative reliability of each signal in domain-specific term matching. Direct term matching was given the second-highest weight (35%) because exact or near-exact lexical overlap is the most unambiguous indicator of a known term, but was ranked below semantic similarity (40%) to allow the system to still recover paraphrased or morphologically varied inputs that a lexical match alone would miss. Word overlap (15%) and length similarity (10%) serve as tie-breaking signals that reduce false positives when two candidate terms have similar semantic embeddings but differ in surface form or length. A cache system is used to store embeddings on disk after they are initially computed to avoid re-computing them on each request.

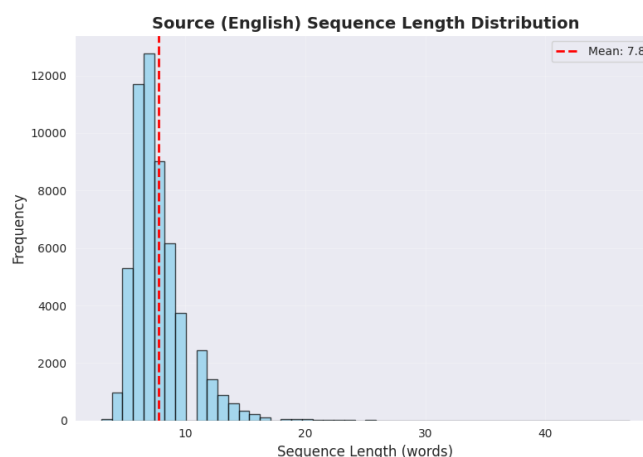


Figure 3. Source (English) Sequence Length Distribution for the English-Spanish Parallel Corpus (Mean: 7.8 Words).

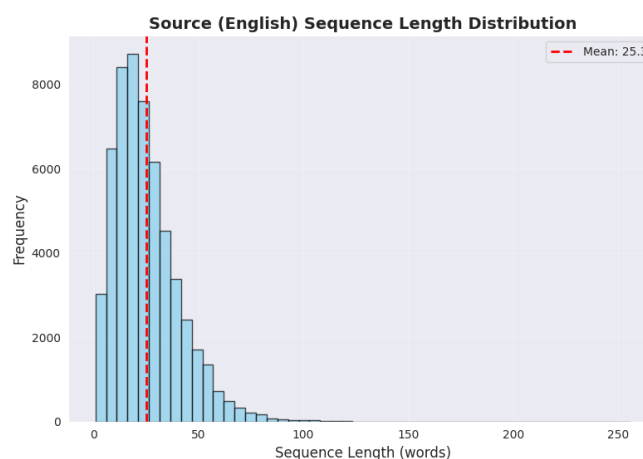


Figure 4. Source (English) Sequence Length Distribution for the English-German Parallel Corpus (Mean: 25.3 Words).

Three separate MarianMT models were fine-tuned from Helsinki-NLP pre-trained checkpoints: opus-mt-en-es, opus-mt-en-fr, and opus-mt-en-de. Training used SentencePiece tokenisation with a maximum length of 128 tokens, a learning rate of 1e-5, and a batch size of 32 on Kaggle T4 GPU instances. The Spanish model was trained for one epoch, since its shorter sentence structures made the model susceptible to overfitting beyond this point. The French and German models were trained for up to five epochs with early stopping based on BLEU score improvements, as the greater sentence complexity in those datasets meant the models could continue learning across multiple epochs without overfitting. The full set of training hyperparameters is summarised in Table 1. All three models used the AdamW optimiser with fused implementation, a linear learning rate schedule, and mixed precision training. The Spanish model additionally used label smoothing (factor 0.1) and a warmup of 500 steps in order to stabilise early training given its single-epoch schedule.

Table 1. Training Hyperparameters for All Three MarianMT Models

Parameter	EN-ES	EN-FR	EN-DE
Learning rate	1e-5	1e-5	1e-5
Train batch size	32	32	32
Eval batch size	128	128	128
Epochs	1	5	5

Optimizer	AdamW (fused)	AdamW (fused)	AdamW (fused)
LR scheduler	Linear	Linear	Linear
Warmup steps	500	—	—
Label smoothing	0.1	—	—
Mixed precision	Yes	Yes	Yes
Random seed	42	42	42

D. System Integration and Deployment

The complete system integrates both models in a unified workflow. User input, whether typed or provided via voice using OpenAI Whisper (whisper-base), is processed by SBERT for term extraction and by MarianMT for translation. Extracted term definitions are translated alongside the input text, so users receive the translated sentence together with contextual explanations for all detected domain-specific terms. Figure 5 illustrates the full system architecture, showing how a voice input passes through Whisper transcription, then through SBERT for term identification and MarianMT for translation, with the final output combining the translated sentence and term-level explanations.

The application was deployed locally as a Flask web application with an HTML/CSS/JavaScript front-end where the user can select the domain, set the strictness, input voice commands, and text-to-speech output via Google Text-to-Speech.

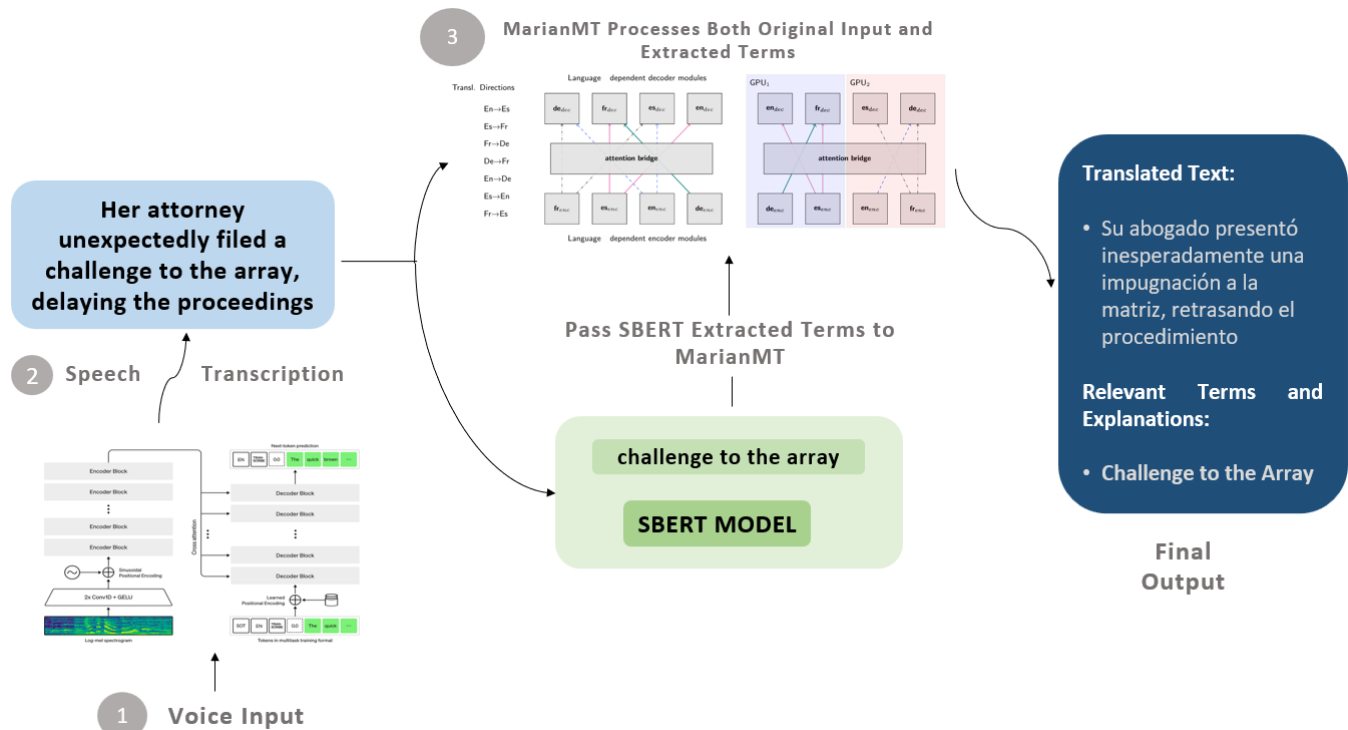


Figure 5. System Architecture Workflow: Voice Input is Transcribed by Whisper, SBERT Extracts Domain-specific Terms, and MarianMT Translates Both the Sentence and Term Definitions to Produce the Final Output.

E. Evaluation Approach

The translation quality was measured using BLEU scores [23], which calculate n-gram overlap between system output and human reference translations on a scale of 0 to 100, where

higher values indicate closer agreement with reference translations. The BLEU score is computed as shown in Formula 1:

$$\text{BLEU} = \text{BP} \cdot \exp(\sum_{n=1}^N w_n \log p_n) \quad (1)$$

where BP is the brevity penalty that discourages translations shorter than the reference, w_n is the uniform weight assigned to each n-gram order (typically $1/N$), p_n is the precision score for n-grams of length n , and N is the maximum n-gram order considered, typically set to 4. The brevity penalty is defined as Formula 2:

$$\text{BP} = \begin{cases} 1, & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r \end{cases} \quad (2)$$

where c is the length of the candidate translation, and r is the length of the reference translation.

The performance of the term extractor was assessed across four test conditions: Normal, which tests exact terms as they are in the corpus; Misspelling, which tests terms with typographic misspellings such as replacing double characters with a single character; Capitalisation, which tests terms with different letter case); and Incomplete, which tests partial term entries that simulate real-world user behaviour where only part of a technical term may be recalled. For each condition, accuracy, precision, recall, and F1-score were calculated.

IV. RESULTS AND DISCUSSION

A. Translation Model Performance

The three MarianMT models achieved the following test BLEU scores: English-Spanish 59.26, English-French 38.76, and English-German 25.68, as summarised in Table 2. The English-Spanish model was trained for one epoch and its validation BLEU climbed steadily from 56.5 at the start of training to a plateau around 59.0 by the end, while validation loss dropped from 1.97 to 1.93, as shown in Figure 6. This score reflects the fact that the Spanish training data contains shorter, structurally simpler sentences. With most sentences between 5 and 10 words long, the model has fewer grammatical patterns to resolve and its output is more likely to align closely with the reference translations word for word.

The English-French model was trained over five epochs. Validation BLEU improved from 34.17 in the first epoch to a peak of 34.50 around epoch 3. As shown in Figure 7, validation loss decreased from 1.137 to 1.114 across the full training run. The lower BLEU scores than Spanish are a direct result of the longer sentences in the French dataset. The average length of the source sentences is 25.3 words, which means that there are many more positions along the source sentence where a generated translation can differ from the reference sentence without changing the meaning. A BLEU score in the mid-to-high 30s for this type of data is considered to be coherent and useful for professional purposes.

Table 2. Bleu Scores for the Three Language Pairs

Language Pair	BLEU Score
English-Spanish	59.26
English-French	38.76
English-German	25.68



Figure 6. Validation Loss for the English-Spanish Model Across Training Steps.

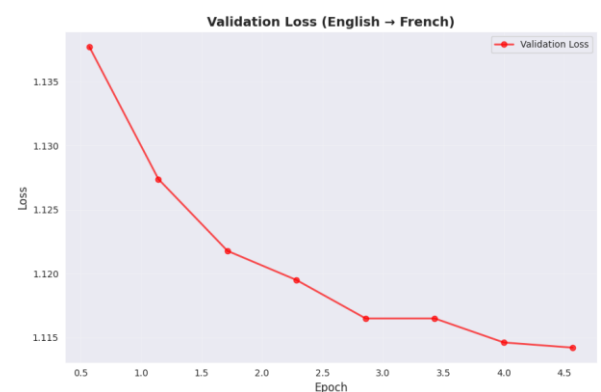


Figure 7. Validation Loss for the English-French Model During Training.

The English-German model reached a peak validation BLEU of 26.57 around epoch 4. The validation loss decreased from 1.426 to 1.401 across training, as shown in Figure 8. German is, however, very challenging because of its flexible word order, its intensive compound word formation, and its complex case system. There is a possibility for a German translation to be grammatically correct, semantically correct, perfectly readable, but still get a low BLEU score due to the different construction of compounds or clause order. The score therefore understates the usefulness of the German model in practice.



Figure 8. Validation Loss for the English-German Model Across Training Epochs

B. Term Extraction Performance

The term extraction system based on SBERT was evaluated in all three domains and the results are shown in Table 3. The embedding approach has been found to be reliable in handling letter case variation as it achieves a perfect accuracy of 100% in all the domains for Normal and Capitalisation test cases. The accuracy for Misspelling cases ranged from 79% in Medical to 94% in Financial. The lower medical figure is due to the tendency of medical terms to be longer and morphologically complicated. If a term is misspelled, for example, "thrombocytopenia", it can be confused with several different terms in the corpus, making its correct identification more difficult. Financial terms are shorter and more standardised, and are still distinguishable even in the face of common typographical errors. In the three domains, the performance of the Incomplete terms ranged between 90% and 92%, meaning that the system can be used to identify a term from a partial input in most cases.

Table 3. Term Extraction Accuracy by Domain and Test Condition

Condition	Legal (%)	Medical (%)	Financial (%)
Normal	100	100	100
Misspelling	82	79	94
Capitalisation	100	100	100

Incomplete	90	92	92
Average	93.0	92.8	96.5

The overall average accuracy was 96.5% for Financial, 93.0% for Legal, and 92.8% for Medical. The small difference between the top and bottom performers (3.7 percentage points) indicates that the SBERT approach generalises well across the domains of varying size and terminological character. The Incomplete condition results are especially important when considering real-world implementation, because professionals in a high-stakes environment are likely to use technical terms informally, and a system that can still recognize the correct term with an imperfect input is substantially more useful than one that requires exact wording.

C. System Interface

The deployed web application offers an accessible interface for all three translation domains. The voice recording interface is represented in Figure 9, and the user is able to see a visualisation of the waveform when recording audio input before submitting it for transcription and translation. Figure 10 shows the financial domain translation interface, which displays a sentence related to an accelerable guarantee being translated into Spanish, and four specialised financial terms translated with similarity scores and bilingual definitions.

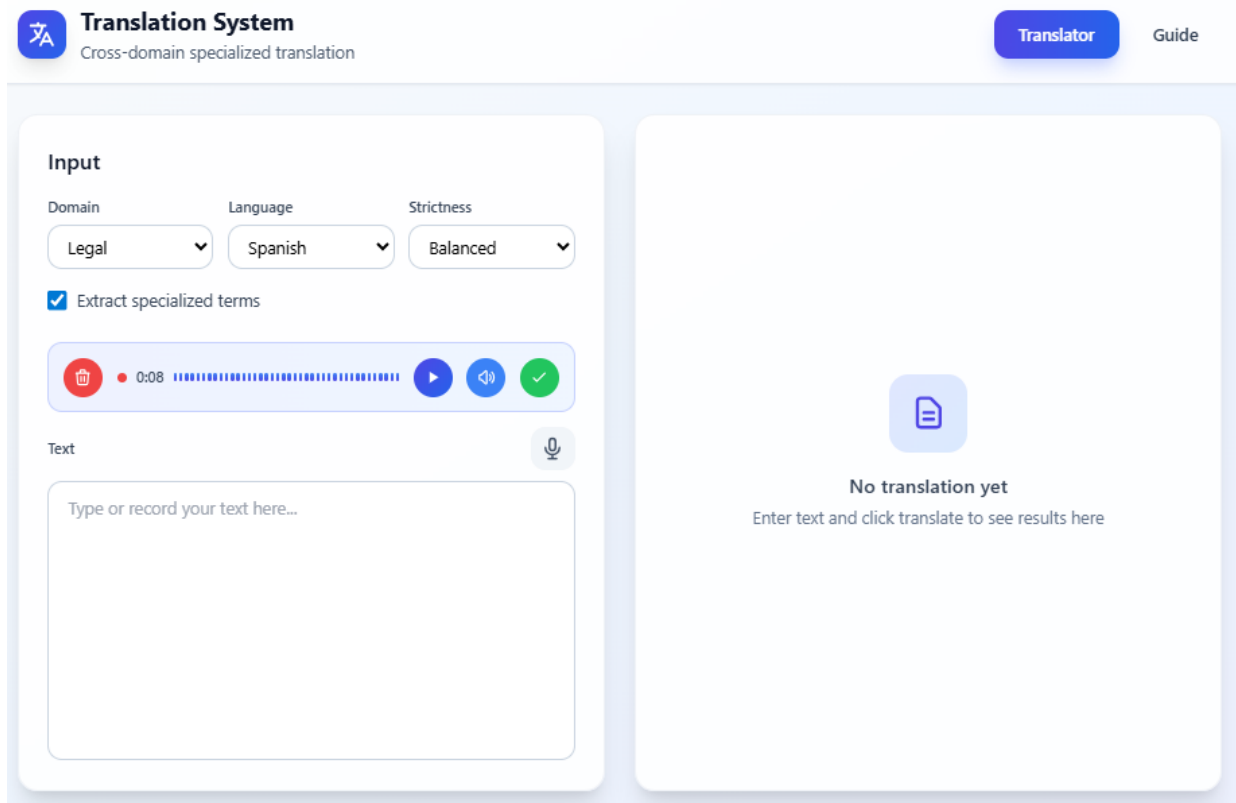


Figure 9. Voice Recording Interface Showing Waveform Visualisation During Audio Input Capture.

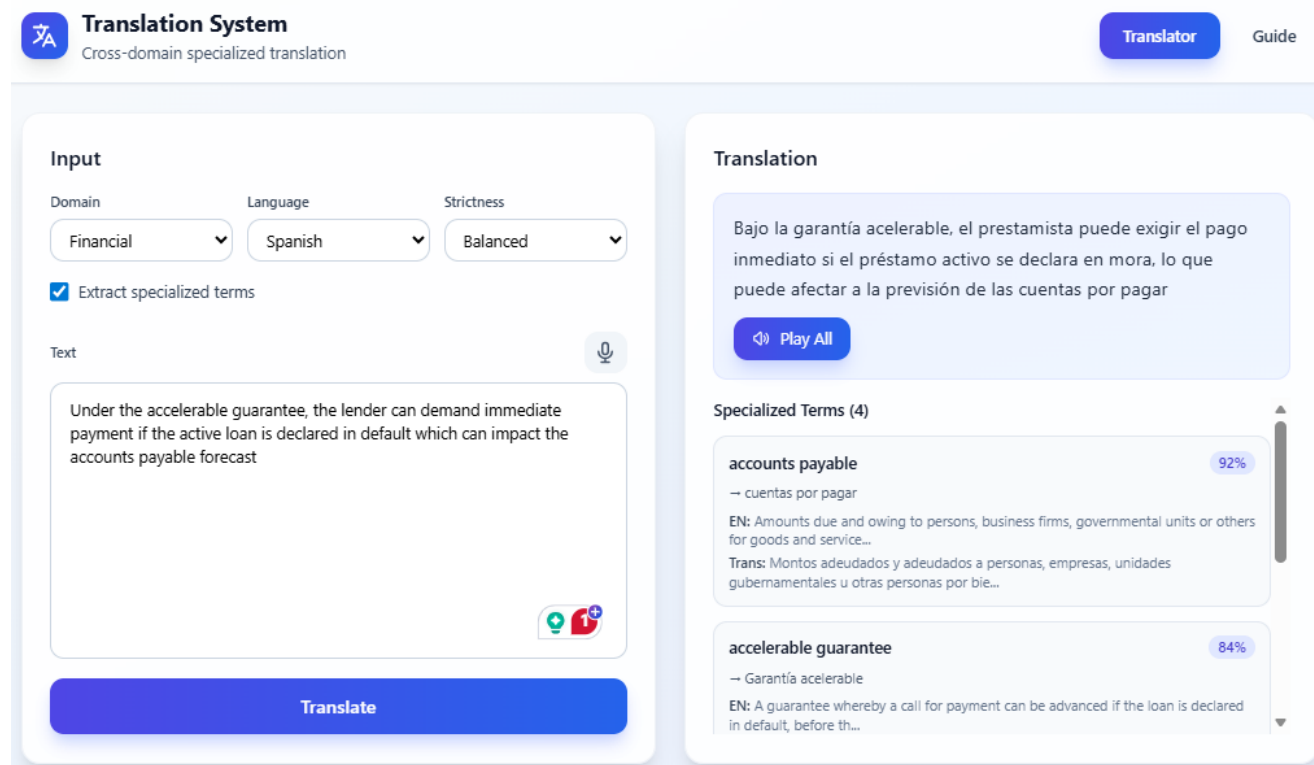


Figure 10. Financial Domain Translation Interface Showing Extracted Specialised Terms With Similarity Scores and Bilingual Definitions.

D. Discussion

The BLEU score differences between the language pairs must be interpreted in conjunction with the length of sentences in the language pairs, and not as a direct comparison of model quality. When sentences are short and structurally simple, as in the Spanish dataset, the model has fewer decision points during generation and the resulting output is more likely to match the reference translation word for word. In the French and German data sets, the sentence length and syntactic variety may lead the model to generate a translation that is semantically correct but not exactly the same as the reference in terms of how it is structured, and BLEU penalises that lack of similarity. The French and German models are therefore more practically adequate than a direct score comparison would suggest.

The lower scores for French and German also reflect the inherent challenge of morphologically rich target languages. German's flexible word order, compound word formation, and case system mean that a grammatically and semantically correct translation can differ significantly from the reference translation at the surface level, causing BLEU to underestimate the translation's actual quality. The validation loss curves for both models (Figures 7 and 8) confirm that learning was genuine and stable rather than stagnant, supporting the view that the scores represent a functional system rather than a failing one.

The hybrid similarity-based method for term extraction is more effective than relying on semantic similarity-based method alone. Using only semantic similarity would increase the risk of false positives in domains where many terms share a conceptual neighbourhood. The weighted sum of the

semantic, lexical and length-based signals enables the system to be more conservative, suitable for professional use where the wrong term explanation can be more harmful than a missed detection.

A comparison with general-purpose commercial systems further contextualises these results. As shown in the introduction, Google Translate and DeepL both failed to produce domain-accurate translations of the legal expression "challenge to the array," offering plausible-sounding but functionally misleading outputs with no contextual explanation. The proposed system, by contrast, not only translates the sentence but identifies the term and retrieves its domain-specific definition in the target language. This qualitative difference is not captured by BLEU alone, which measures surface-level n-gram overlap with reference translations rather than domain accuracy or explanatory adequacy. A system that scores lower on BLEU while providing correct domain explanations may be substantially more useful in practice than a higher-scoring system that produces fluent but contextually empty output.

One significant drawback is the fact that the system is completely reliant on its domain corpora. If a term is not in the legal, medical, or financial database, then it cannot be identified or explained, even if it is relevant to the user input. This is a limitation of retrieval-based methods and suggests that expanding the corpus is the most obvious way to increase coverage. Another drawback is that the Spanish training data comes from conversational sources with shorter sentences, which could lead to poorer results on formal legal or medical

texts in Spanish, which tend to be longer and more complex in structure.

V. CONCLUSION

This study demonstrated that Sentence-BERT for domain-specific term extraction and fine-tuned MarianMT for translation can fill a practical need in current translation tools, which is the lack of contextual explanations for technical terminology. For English-Spanish, English-French and English-German translation, the system attained BLEU scores of 59.26, 38.76, and 25.68, respectively, with the training curves being stable in all three. The accuracy of term extraction was 92.8 to 96.5% in all the legal, medical, and financial domains, with performance on incomplete and misspelled inputs remaining above 79% in all conditions.

The system has two main limitations that should be acknowledged. First, term detection is fully dependent on the domain corpora: if a term does not appear in the legal, medical, or financial database, the system cannot identify or explain it, regardless of its relevance to the user input. Second, the Spanish training data draws from conversational sources with short sentences, which may reduce performance on formal professional Spanish texts that tend to be longer and syntactically more complex. Addressing these limitations is the primary direction for future work.

Future research should aim to create more diverse training corpora for each language pair, covering both formal and informal registers, for better performance in professional documents. A more flexible system for real-world deployment would be to use adaptive strictness thresholds for term extraction that adjust dynamically according to the domain, term length and input context. It would also save storage and maintenance costs if multi-domain models could be developed that could handle mixed content without having to maintain separate per-domain databases. In terms of evaluation, metrics that weight terminology accuracy more directly than n-gram overlap would give a more faithful account of system performance in specialised professional contexts. Extending the system to additional languages, including Indonesian and other low-resource languages with significant professional translation needs, would also increase the scope and impact of the approach.

The results show that using semantic term extraction in conjunction with neural machine translation produces more informative output than translation alone. In professional domains such as medicine, law, and finance, knowing the meaning of a translated word is sometimes as important as the translation, and the hybrid method described here provides a practical way to provide both.

REFERENCES

- [1] G. R. Vlasian, O. D. Pastukhova, and V. A. Shusharina, "On hedging and mitigation in the speeches of Boris Johnson," in *Proc. European Conf. Social & Behavioural Sciences*, 2020, pp. 185-194.
- [2] W. R. Leben, "Languages of the world," *Oxford Research Encyclopedia of Linguistics*, 2016. doi: 10.1093/acrefore/9780199384655.013.349
- [3] P. Koehn, *Neural Machine Translation*. Cambridge: Cambridge University Press, 2020.
- [4] S. Sharma et al., "Machine translation systems based on classical-statistical-deep-learning approaches," *Electronics*, vol. 12, no. 7, p. 1716, 2023.
- [5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.
- [6] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.
- [7] M. Junczys-Dowmunt et al., "Marian: Fast neural machine translation in C++," *arXiv:1804.00344*, 2018.
- [8] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv:1910.10683*, 2019.
- [9] P. Naveen and P. Trojovský, "Overview and challenges of machine translation for contextually appropriate translations," *iScience*, vol. 27, no. 10, p. 110878, 2024.
- [10] P. Sebo and S. De Lucia, "Performance of machine translators in translating French medical research abstracts to English," *PLoS ONE*, vol. 19, no. 2, e0297183, 2024.
- [11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," *arXiv:1908.10084*, 2019.
- [12] M. Müller, A. Rios, and R. Sennrich, "Domain robustness in neural machine translation," *arXiv:1911.03109*, 2019.
- [13] S. Agrawal, C. Zhou, M. Lewis, L. Zettlemoyer, and M. Ghazvininejad, "In-context examples selection for machine translation," *arXiv:2212.02437*, 2022.
- [14] U. Khandelwal, A. Fan, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Nearest neighbor machine translation," *arXiv:2010.00710*, 2020.
- [15] E. A. Alghamdi, J. Zakraoui, and F. A. Abanmy, "Domain adaptation for Arabic machine translation: Financial texts as a case study," *Applied Sciences*, vol. 14, no. 16, p. 7088, 2024.
- [16] N. Limsopatham and N. Collier, "Normalising medical concepts in social media texts by learning semantic representation," in *Proc. ACL*, 2016.
- [17] J. Zeng et al., "Multi-domain neural machine translation with word-level domain context discrimination," in *Proc. EMNLP*, 2018.
- [18] S. Poudel, B. K. Bal, and P. Acharya, "Bidirectional English-Nepali machine translation system for legal domain," in *Proc. SIGUL, ACL Anthology*, 2024.
- [19] R. Pierce, W. Charlton, and C. Murphey, Eds., *Handbook of Legal Terminology*, 4th ed. University, MS, USA: Mississippi Judicial College, University of Mississippi, 2020.
- [20] Merriam-Webster, *Merriam-Webster's Medical Dictionary*, 3rd ed. Merriam-Webster, 2008.

- [21] International Bank for Reconstruction and Development/The World Bank, World Bank Glossary of Finance and Debt. The World Bank, 1991.
- [22] Financial Glossary. San Francisco, CA, USA: City College of San Francisco, [Online]. Available: <https://www.ccsf.edu/sites/default/files/2024/document/financial-glossary.pdf>. [Accessed: May 10, 2026]
- [23] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. ACL, 2002.

