Flipper: An Extended Document-Level Financial Dataset for Training and Evaluation with Annotated Discourse Phenomena

Mariam Nakhlé[1,2], **Rachel Atherly**[1], **Gabriela González Sáez**[1], **Marco Dinarelli**[1], **Raheel Qader**[2], **Hervé Blanchon**[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP , LIG, 38000 Grenoble, France
(2) Dragon LLM, 75008, Paris, France
contact: mariam.nakhle@univ-grenoble-alpes.fr

Abstract

We present a new resource for Machine Translation (MT), namely a training and evaluation dataset containing parallel documents issued from authentic data in the financial domain. We cover five language pairs: English-French, English-Spanish, English-German, English-Italian and French-Spanish. The total number of parallel documents is 122k and the number of tokens is 118M (source and target combined). MT has improved greatly in recent years, but certain phenomena still cause errors, particularly when context spans beyond a single sentence. Errors can lead to mistranslated pronouns, incorrect gender or number agreement, and inconsistent terminology, which can be especially problematic in high-stakes domains like finance. We therefore construct the dataset at document level (rather than sentence-level) and also produce fine-grained annotations of context-sensitive phenomena. The annotation was performed using preexisting tools and custom scripts. Thus the process can be replicated on different parallel data from other domains. The annotated phenomena are: formality, gender, terminology consistency, verb form and sentence reordering. This aims to improve document-level evaluation of MT models by enabling evaluation solely on texts containing a particular phenomenon of interest. Our primary contribution is the creation and public release of Flipper, a multilingual document-level parallel dataset in the financial domain, designed to support both training and targeted evaluation of context-sensitive machine translation.
Keywords: Financial NLP, Machine Translation, Parallel data, Annotation

Lang. Pair	Documents	Sentences
ENDE	34,438	1,268,547
ENES	12,059	286,376
ENFR	47,327	1,399,169
ENIT	14,449	330,128
FRES	14,588	325,318
TOTAL	**122,861**	**3,609,538**
	Src. tokens	**Trg. tokens**
ENDE	20,351,372	18,347,846
ENES	4,942,426	5,403,083
ENFR	22,119,921	23,811,471
ENIT	5,038,081	5,179,604
FRES	6,206,158	6,253,966
TOTAL	**58,657,958**	**58,995,970**

Table 1: Statistics of the full dataset. For every language pair, 2000 documents are reserved for the evaluation split.

1. Introduction

While Context-Aware Neural Machine Translation (CA-NMT) has received considerable attention in recent years (Toral et al., 2018; Läubli et al., 2020), with numerous works focusing on architectural improvements and modelling strategies, advances in context-aware evaluation have not progressed at the same pace. Many studies still report quantitative results based on sentence-level metrics such as BLEU or COMET (Papineni et al., 2002; Rei et al., 2020), which are not designed to capture document-level phenomena. In parallel, several works rely on ad-hoc evaluation datasets, such as contrastive test suites (Bawden et al., 2018; Müller et al., 2018), that target specific discourse phenomena but do not reflect realistic document-level translation settings.

This imbalance between modelling and evaluation is particularly evident in specialised domains such as finance. Despite the growing importance of domain-adapted MT, very few document-level datasets have been released for the financial domain, and even fewer explicitly target context-sensitive phenomena.

The financial domain constitutes a highly relevant and challenging use case for context-aware MT. It is characterised by specialised terminology, strict regulatory requirements, and a high demand for fast and reliable translations. Providers of financial products are legally required to supply translated versions of documentation in every country where the product is commercialised, creating a strong need for high-quality, in-domain translation systems.

Financial documents exhibit several properties that make document-level evaluation especially crucial. Terminology must remain consistent throughout a document, and the correct translation of a term often depends on definitions introduced earlier in the text—particularly in legal documents, where key entities are explicitly defined at the beginning and must be referred to consistently thereafter. In

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 78–86

addition, numerical consistency is essential: financial reports frequently contain tables with monetary amounts, and inconsistent formatting (e.g., 5 million USD, 5M USD, or $5M) within or across tables is unacceptable. These characteristics highlight the limitations of sentence-level evaluation and underline the need for document-level resources tailored to this domain.

To address this gap, we propose Flipper[1], a new document-level parallel dataset for the financial domain designed for both training and evaluation of CA-NMT systems. We collect a large number of publicly available parallel documents issued by asset management firms. The documents, originally in PDF format, are primarily legal and marketing materials, including annual reports, Key Information Documents, marketing factsheets, manager commentaries, and ESG or sustainability reports.

To convert these documents into parallel data suitable for MT training, we adopt a pipeline inspired by Nakhlé et al. (2025). As in their work, we extract texts from PDFs and align full documents or sections rather than individual sentences, enabling the preservation of higher-level discourse structure and allowing sentence reordering within sections. However, Flipper substantially extends and improves upon this approach. We introduce a more robust alignment method and perform extensive deduplication of near-duplicate content through custom preprocessing steps, resulting in more diverse data, but still a much larger dataset. Beyond data construction, we perform careful annotation of context-sensitive phenomena that are annotated using an inline XML format. The annotation was performed using preexisting annotation tools and custom-made scripts. The annotated phenomena are: formality, gender, terminology consistency, verb form and sentence reordering.

Unlike the DOLFIN dataset, Flipper includes both training and evaluation splits, which makes it a valuable resource for model training. 2000 documents are reserved for the evaluation set, we deliberately select sections with a strong presence of discourse- and context-dependent phenomena. This design positions our evaluation split between a standard test set and a targeted test suite: it remains authentic and document-based, yet it increases the density of challenging phenomena, making evaluation more informative.

Flipper covers five language pairs: English–Italian (En–It), English–Spanish (En–Es), English–German (En–De), English–French (En–Fr), and French–Spanish (Fr–Es), the latter being the only non-English pair. As stated in Table 1, the total number of parallel sections is 122k and the number of tokens is 118M (source and target tokens combined).

The dataset can be used to train or fine-tune document-level MT systems or Large Language Models (LLMs) and to conduct targeted evaluation of specific context-sensitive phenomena through its fine-grained inline annotations. To the best of our knowledge, Flipper is the first publicly available parallel dataset in the financial domain that simultaneously satisfies three key criteria: (1) document-level structure, (2) domain-specific coverage of authentic financial texts, and (3) suitability for both training and targeted evaluation. Our contributions can be summarized as following:

- The creation and public release of a document-level parallel dataset for the financial domain;
- The provision of a dedicated training split tailored to document-level MT;
- The targeted selection of evaluation data enriched with discourse- and context-sensitive phenomena;
- Fine-grained inline annotation of multiple context-sensitive phenomena, including formality, gender, terminology consistency, verb form, and sentence reordering.

2. Related work

Natural Language Processing (NLP) for finance. The use of NLP techniques across various applications has grown significantly in recent years across a broad range of domains, and finance is no exception. Some of the well-studied tasks in this field include named entity recognition (Salinas Alvarado et al., 2015), question answering (Chen et al., 2021; Maia et al., 2018), sentiment analysis (Malo et al., 2014), and topic modeling (Jehnen et al., 2025).

Several finance-oriented language models have also been developed. These include encoder-only models such as FinBERT (Araci, 2019), as well as decoder-only models such as BloombergGPT (Wu et al., 2023), FinMA (Xie et al., 2023), FinGPT (Wang et al., 2023), and LLM Pro Finance (Caillaut et al., 2025). These models can be applied to a variety of downstream tasks, including Machine Translation, which is the focus of this work.

Machine Translation for finance. Although Machine Translation (MT) is one of the core tasks in NLP, resources tailored to the financial domain remain scarce. With regard to financial-domain resources, the diachronic banking magazine collections (Volk et al., 2016) provide relevant material; however, the data is available only at the sentence level. Nakhlé et al. (2025) present a document-level parallel dataset for the financial domain, but it is limited to evaluation data. More generally, several document-level datasets have been proposed for

[1] https://huggingface.co/datasets/DragonLLM/Flipper

training purposes (Koehn, 2005; Tiedemann, 2012; Cettolo et al., 2012; Lison and Tiedemann, 2016; Wicks et al., 2024). As for evaluation data, the majority of available benchmarks remain sentence-level, such as the datasets released annually within the WMT conference (Kocmi et al., 2025). Some test sets do include document boundaries; however, even in these cases, the primary unit remains the sentence, meaning that individual sentences are aligned within documents.[2] Evaluation typically measures the overall quality of a model by feeding translations to the metric one sentence at a time. By design, this excludes looser translation strategies, such as sentence reordering or splitting. In contrast, our approach aims to accommodate such phenomena, which is why the main unit of our proposed dataset is the document rather than the sentence.

Another line of work in context-aware evaluation involves targeted test suites that focus on linguistic phenomena requiring context beyond a single sentence for correct evaluation. These resources are often constructed manually and are designed to probe systems with respect to specific discourse-level challenges. While they provide valuable insights into particular phenomena, they remain limited in scope, language coverage, and applicability to overall translation quality.

In this work, we aim to address the gap in available resources by providing parallel data that meet three criteria: (1) document-level structure, (2) coverage of the financial domain, and (3) suitability for both training and evaluation. Our dataset, Flipper, enables the training and evaluation of MT models on challenging texts drawn from authentic financial documents containing context-sensitive phenomena, thereby combining general quality assessment with targeted evaluation.

3. Dataset collection

Our dataset collection procedure is made of the following processing steps: 1) PDF-to-text (Markdown) extraction, 2) noisy data filtering, 3) alignment of sections within documents, 4) near-duplicate removal, and 5) context-sensitive phenomena annotation. This pipeline is inspired by the one presented in Nakhlé et al. (2025), as we similarly worked with financial PDF documents and aimed to produce a document-level dataset. However, we introduce improvements in certain aspects, particularly in the alignment approach, which is crucial for parallel data. We also modified the deduplication process, as the data still appeared to be highly repetitive. In contrast to the cited previous work, we do not perform quality estimation filtering, as we believe it introduces bias into the data curation process. Finally, we produced fine-grained inline annotations of context-sensitive phenomena.

Alignment. We used a different alignment approach, as the original method was rather naive (two sections were considered aligned if their first and last sentences were aligned). In our approach, we use the same aligner LASER (Schwenk and Douze, 2017), but we modified the decision logic for determining whether two documents are aligned.

We compute LASER alignment scores between each source sentence and a window of size m in the target document. For a source sentence with index i, the target window ranges from `i - m/2` to `i + m/2`. We then compute an average score for the entire section based on the scores of all source sentences, taking the maximum score for each source sentence.

To determine an acceptability threshold, we manually inspected 1,200 sections. Many collected sections were challenging cases: although they appeared similar, they were not faithful translations of one another. A threshold of 0.80 was selected as the best balance between permissiveness and strictness, while retaining the maximum number of correctly aligned sections.

Deduplication. Another limitation of the dataset described by Nakhlé et al. (2025) is its repetitiveness, which is directly related to the financial domain. Companies are legally required to publish certain documents periodically; for efficiency, these documents often contain repeated sections. Additionally, some documents—such as Key Information Documents—must present information in a standardized format. As a result, texts describing different funds are often identical, differing only in numerical values.

To address this issue, we applied stricter deduplication using the MinHash algorithm implemented in the *text-dedup* library by Mou et al. (2023), complemented by a preprocessing step. Specifically, we tokenized the texts and masked numerical values so that texts differing only in numbers would still be identified as duplicates and removed. We then set the deduplication threshold to 0.5. For this step, we considered merged source and target documents.

4. Annotation

In order to enrich the dataset with more nuanced information, we performed an annotation of context-sensitive phenomena. These are linguistic phenomena that require extra-sentential information to be correctly interpreted and translated, and they present a particular challenge for Machine Translation, especially when models operate on isolated sentences. This annotation enables future users

[2] To the best of our knowledge, the WMT test sets released in 2025 are the first ones in which the document is treated as the main unit.

of the dataset to target specific phenomena by extracting the relevant sections and analysing model performance accordingly.

To carry out the annotation, we employed several pre-existing tools designed for this purpose. Formality, plurality, and grammatical gender are among the more common context-sensitive phenomena and were annotated using existing tools developed for parallel data annotation. To further diversify the range of annotated phenomena, we developed custom tools to address two additional aspects: terminology consistency, which ensures that terms are translated consistently throughout a document, and high-level information reordering, which captures changes in sentence order during translation.

4.1. Annotation format

The annotations are provided in an inline XML format. This standardises the notation across different tools and annotation methods, enabling multiple phenomena to be marked within the same section.

The main tag used is `<annotation>`, which includes a `tool` attribute indicating the source of the annotation. The `phenomenon` attribute specifies the type of context-sensitive phenomenon being marked. Depending on the phenomenon, additional attributes provide more detailed information. Below are two examples of annotations, illustrating the formality phenomenon and the terminology consistency phenomenon, respectively.

> Before buying or switching Units, <annotation tool="ctxpro" rule="NOM.FORM+PLUR" phenomenon="formality"> you </annotation> should read the relevant KIID.

> The absolute and/or relative returns shown in the performance attribution section of this document (hereafter <annotation tool="custom" phenomenon="terminology consistency" id="def_1" refersTo="performance attribution section"> Reporting </annotation>) may differ from the returns in other statements or reports provided due to the use of different methodologies. For official use, only the official statements and reports should be used and not the figures provided in this <annotation tool="custom" phenomenon="terminology consistency" id="ref_1" refersTo= "def_1"> Reporting </annotation>.

4.2. Annotations using preexisting tools

4.2.1. CTXPRO

CTXPRO (Wicks and Post, 2023) is an automatic annotation tool that identifies phenomena that require contextual information to be translated correctly. The identification process is rule-based and relies on linguistic information specific to the language pair. The pipeline takes sentence-aligned parallel data as input and outputs information indicating which segments contain ambiguities.

CTXPRO was run on all language pairs except French–Spanish, as the tool requires at least one of the languages to be English. It was able to tag phenomena related to gender and formality ambiguities. The formality phenomenon arises when one language uses an ambiguous pronoun whose translation into the target language depends on the level of formality. For example, the English pronoun “you“ translates into French as “tu“ (informal) or “vous“ (formal). The gender phenomenon occurs with pronominal anaphora, where correct translation of a pronoun requires access to its antecedent. For instance, the English pronoun “it“ may translate into French as “il“ (masculine) or “elle“ (feminine), depending on the gender of the antecedent in the translation.

In the annotation, we include the rule as provided by the tool via the `rule` attribute. An example of a rule is NOM.FORM.SING. We use the middle element of the notation to indicate the phenomenon category, in this case formality.[3]

4.2.2. MuDA

The second pre-existing annotation tool employed was the Multilingual Discourse-Aware (MuDA) benchmark (Fernandes et al., 2023). Similarly to CTXPRO, MuDA identifies and annotates phenomena that require context spanning multiple sentences or longer stretches of text. The tool shares the same limitation of requiring English to be part of the language pair. On our data, the phenomena identified by this tool are verb form and formality. The verb form phenomenon arises when the choice of a verbal form (such as tense or mood) depends on extra-sentential context, as is the case in text passages containing a succession of verbs in the imperfect tense.

4.3. Hand crafted annotations

4.3.1. Terminology consistency

Financial documents contain numerous terms that are defined or redefined at a specific point in the

[3]We refer the reader to the original paper for a detailed description of the rules.

text. For example, in the sentence “the total risk is measured and checked using the relative value at risk (hereinafter ‘relative VaR’) method,“ the term “relative VaR“ must remain unchanged throughout the rest of the document during translation. A similar situation arises with abbreviated company names. In our financial documents, it is common for the full name or legal designation of a financial fund to be stated at the beginning and subsequently referred to as “the Fund” throughout the remainder of the text. This creates terms that require consistent and accurate translation, even though their definition appears only once, several sentences or even paragraphs earlier. This phenomenon is particularly critical in financial and legal contexts, where the acceptable margin of error is significantly lower than in other domains.

To identify instances of terminology consistency, a semi-manual program was designed and implemented. First, regular expressions were used to search the dataset for the trigger words “hereafter” and “hereinafter,” which commonly signal the introduction of a definition, as illustrated in the example above. The term introduced as the new definition (referred to as the “alias”) was then identified using an additional set of regular expression patterns. However, due to numerous edge cases and variations in wording, identifying the term being redefined (the “head”) required more manual effort. While the pattern could detect potential instances of terminology consistency, accurately annotating the details required further validation. Therefore, an interactive interface was developed that used the regular expressions to display the predicted “alias” and “head” pairs individually for review. Using this approach, the phenomenon was annotated with human oversight to approve, reject, or edit each tag. This method enabled efficient annotation while increasing confidence that the final dataset did not contain erroneous tags resulting from unique edge cases.

This phenomenon includes attributes `id` and `refersTo` in the annotation tags. The `refersTo` attribute links a term to its corresponding definition. The `id` attribute can take the form `def_x` or `ref_x`, depending on whether the tag marks the definition of a terminology rule or a subsequent reference to it. The attribute `id="def_x"` appears within hereafter declarations, while `refersTo` contains the word or phrase introduced by the definition. If the tag marks a later reference to an existing definition, `refersTo` contains only the corresponding definition ID. This structure makes it possible to count how many references a given terminology rule has. Definition IDs are global (e.g., there is only one `def_1`), whereas reference IDs are local; for instance, `ref_1` denotes the first reference associated with each `def_x`.

4.3.2. Sentence reordering

During translation, a certain degree of high-level information reordering may occur. It is not always possible to translate sentences word for word or to process documents strictly sentence by sentence. In some cases, sentences must be reordered, split, or merged to produce a more natural and fluent text in the target language. This poses a particular challenge for MT evaluation, as the source and target texts may not preserve sentence-to-sentence alignment, while evaluation metrics typically process sentences individually. For this reason, we sought to annotate this phenomenon in order to enable targeted evaluation and to analyze the extent to which it occurs in the translation of financial documents.

To annotate sentence reordering, we first preprocessed the texts by removing tables and other heavy *Markdown* formatting. The sentences were then encoded using a Sentence-BERT model (Reimers and Gurevych, 2019)[4]. We computed a cosine similarity matrix between source and target sentences to identify the best match for each sentence. If the best match for a given source sentence had a different index in the target text than in the source, it was flagged as a potential reordering. This method detects both direct swaps and broader shifts in sentence order. In several sections, for example, the first sentence is split in translation, shifting all subsequent sentence indices by one throughout the section.

The annotation tag for this phenomenon includes the attributes `from` and `to`, which indicate the original and matched sentence indices, respectively. The `from` attribute refers to the index of the source sentence and, by extension, to the expected index of its corresponding translation. The `to` attribute indicates the index of the target sentence identified as the best match. For example, `from=2 to=3` means that source sentence 2 has a higher cosine similarity with target sentence 3 than with target sentence 2, suggesting that a reordering has occurred.

However, manual analysis of the annotations revealed that this method is insufficient for reliably detecting sentence reordering. In many cases, shifts in sentence indices were caused by errors introduced by the sentence splitter or by minor misalignments between source and target texts. For instance, if the first sentence is omitted in translation, all subsequent sentence indices are shifted. Although we retain the existing annotations, further investigation of this phenomenon is left to future work.

[4]HuggingFace identifier: sentence-transformers/distiluse-base-multilingual-cased-v1

4.4. Annotation statistics

Table 2 presents the results of the annotation over the full dataset. The highest number of annotations is observed for the English–French language pair, indicating that the tools are most effective for this high-resource pair. A notable limitation of the annotation process concerns the only non-English-centric language pair, French–Spanish, for which none of the pre-existing tools were suitable. The most frequently annotated phenomenon is verb form, as identified by the MuDA tool, followed by sentence reordering and formality. Manual inspection showed that some of the annotations point to phenomena that are in fact translatable even without context (despite the tools being specifically designed for this goal) and we suspect that there are context-sensitive phenomena that were missed by the tools and exist in the dataset without an annotation. This imbalance highlights the inherent difficulty of constructing targeted test sets from authentic documents that present specific context-sensitive translation challenges.

5. The training and evaluation splits

Finally, the dataset was divided into training and test splits. The test set contains ten thousand sections, with two thousand sections per language pair. These sections were selected based on the quantity and variety of annotation tags in order to maximize phenomenon diversity and thereby create a more challenging evaluation set, with the aim of reserving the most complex sections for the test set. The training set consists of all remaining sections, with a target size of at least ten thousand aligned sections per language pair.

6. Experiments

6.1. Setup

Next, we conducted an experiment by fine-tuning the 1B-parameter Gemma 3 Instruct model (hereafter gemma-3-1b-it) (Team, 2025). We selected this model due to its broad language coverage, and because our preliminary experiments indicated strong performance on machine translation.

We fine-tuned the model on the newly constructed Flipper dataset using full-parameter supervised fine-tuning implemented with the TRL (Transformer Reinforcement Learning) library. We use the chat template applicable for this model and we add the following prompt via the user content: `Translate the following paragraph from {source_language} to {target_language}.`
`n Do not add anything else at all.{source_text}.` The assistant reply is the translation alone.

The loss was computed over the completion tokens only; in our setup, these correspond to the translated text. Training is conducted for 1 epoch with a batch size of 8, learning rate 1e-4, linear scheduler, and paged AdamW 8-bit optimizer. We use bfloat16 precision with gradient checkpointing enabled and a maximum sequence length of 1024 tokens. Training is performed on a single GPU with seed 42.

6.2. Evaluation

For evaluation, we translated the evaluation split of Flipper and compared the fine-tuned model against the base gemma-3-1b-it model in a zero-shot setting. We report results using the `wmt22-comet-da` metric (Rei et al., 2022).

Table 3 presents the results per language pair, while Table 4 details the results per phenomenon. As shown, fine-tuning improves translation quality for three out of the five language pairs. The largest gain is observed for English–Spanish, followed by English–French and French–Spanish. In contrast, English–German and, to a lesser extent, English–Italian show a slight degradation compared to the base gemma-3-1b-it model.

At the phenomenon level, four phenomena out of five show gains (Formality, Terminology Consistency, Verb Form, and Reordering), with the largest gain observed for Formality and in Verb Form. These gains suggest that the additional training particularly benefits controlled linguistic phenomena. In contrast, performance on Gender decreases, indicating that improvements are not uniformly distributed across all the phenomena.

LLMs typically benefit from exposure to diverse tasks and domains, which help boost performance on the target task (Alves et al., 2024). This may explain the slight degradation observed. Fine-tuning on a narrower dataset or a single task can lead to performance gains in some directions while slightly degrading others, which is why usually the post training phase includes a mix of tasks. A more comprehensive fine-tuning strategy involving multiple tasks and datasets could yield more uniform improvements across languages and phenomena. However, conducting a large-scale, multi-task optimization was beyond the scope of this work. The primary objective of this paper is to introduce the new dataset and demonstrate its usability through a focused fine-tuning experiment, rather than to exhaustively optimize model performance.

Phenomenon	En-Fr	En-Es	En-It	Fr-Es	En-De	Totals
Verb form	88 723	2 332	551	0	0	91 606
Gender	2 203	30	22	0	609	2 864
Formality	6 628	360	481	0	7 513	14 982
Term. consistency	1 360	284	409	0	1 340	3 393
Sent. reordering	6 833	2 512	2 726	2 358	6 611	21 040
Totals	105 747	5 518	4 189	2 358	16 073	133 885

Table 2: Results of the annotation. We report the number of annotation tags for all phenomena (meaning multiple tags can be present in one document), except for the sentence reordering where we report the number of annotated documents, since one shift can cause all the subsequent sentences to be annotated as reordered.

Lang	gemma-3-1b-it	Fine-tuned
En-De	69,77	67,81
En-Es	77,90	80,25
En-Fr	75,06	75,91
En-It	74,30	73,73
Fr-Es	78,19	78,87

Table 3: Comet scores per language pair.

Phenomenon	gemma-3-1b-it	Fine-tuned
Formality	59,22	61,16
Gender	77,33	75,62
Term. consist.	74,69	75,23
Verb form	74,07	75,86
Reordering	67,00	67,64

Table 4: Comet scores per phenomenon.

7. Conclusion

We presented Flipper, a multilingual, document-level parallel dataset for the financial domain, designed to support both training and evaluation of context-aware machine translation. The dataset covers five language pairs and contains 122k parallel sections with 118M tokens. Our work makes several key contributions: We improved data processing pipeline by improving the section-level alignment, and deduplication process, enhancing the quality and diversity of parallel data compared to prior work. We also added inline fine-grained annotation of context-sensitive phenomena, including formality, gender, terminology consistency, verb form, and sentence reordering, enabling targeted evaluation of challenging translation issues. The dataset has a training and evaluation utility since it includes a dedicated training split and a carefully selected evaluation set enriched with discourse- and context-dependent phenomena, making it suitable for both model training and more informative evaluation. By combining authentic financial documents, document-level structure, and rich annotations, Flipper offers a valuable resource for advancing document-level MT in high-stakes, domain-specific settings.

8. Limitations

Text extraction from PDFs remains a bottleneck and can introduce errors that propagate throughout the processing and annotation pipeline. The French–Spanish pair was particularly affected by tool dependencies on English, highlighting the need for further work on annotations, as existing tools proved largely incompatible. Additionally, our approach may misinterpret sentence reordering due to errors from sentence splitting or minor source–target misalignments, which were not fully addressed in this study. These limitations may be subject of future work.

9. References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Gaëtan Caillaut, Raheel Qader, Jingshu Liu, Mariam Nakhlé, Arezki Sadoune, Massinissa Ahmim, and Jean-Gabriel Barthelemy. 2025. The

llm pro finance suite: Multilingual large language models for financial applications.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Simon Jehnen, Joaquín Ordieres-Meré, and Javier Villalba-Díez. 2025. Fintextsim: Enhancing financial text analysis with bertopic.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the American Society for Information Science and Technology*.

Chenghao Mou, Chris Ha, Kenneth Enevoldsen, and Peiyuan Liu. 2023. Chenghaomou/text-dedup: Reference snapshot.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. *arXiv preprint arXiv:1810.02268*.

Mariam Nakhlé, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, and Hervé Blanchon. 2025. DOLFIN - document-level financial test-set for machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5544–5556, Albuquerque, New Mexico. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu

Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*.

Martin Volk, Chantal Amrhein, Noëmi Aepli, and Phillip Ströbel. 2016. Building a parallel corpus on the world's oldest banking magazine.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets.

Rachel Wicks and Matt Post. 2023. Identifying context-dependent translations for evaluation set production. In *Proceedings of the Eighth Conference on Machine Translation*, pages 452–467, Singapore. Association for Computational Linguistics.

Rachel Wicks, Matt Post, and Philipp Koehn. 2024. Recovering document annotations for sentence-level bitext. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9876–9890, Bangkok, Thailand. Association for Computational Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance.