Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3595–3602
Marseille, 11–16 May 2020

SEDAR: a Large Scale French-English Financial Domain Parallel Corpus

Abbas Ghaddar, Philippe Langlais
RALI-DIRO
Montreal, Canada
abbas.ghaddar@umontreal.ca, felipe@iro.umontreal.ca

Abstract

This paper describes the acquisition, preprocessing and characteristics of SEDAR, a large scale English-French parallel corpus for the financial domain. Our extensive experiments on machine translation show that SEDAR is essential to obtain good performance on finance. We observe a large gain in performance of machine translation systems trained on SEDAR when tested on finance, which makes SEDAR suitable to study domain adaptation for neural machine translation. The first release of the corpus comprises 8.6 million high quality sentence pairs that is publicly available for research at `https://github.com/autorite/sedar-bitext`.

Keywords: Parallel Corpus, Financial Data, Machine Translation

1. Introduction

Neural machine translation (NMT) (Cho et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017; Edunov et al., 2018) has become the most popular approach in recent years for machine translation. These models require the use of large scale parallel corpora to train millions of internal parameters (Ott et al., 2018). The quantity and quality of training data is crucial for systems performances, as well as train/test domain matching. Koehn and Knowles (2017) show that NMT performs poorly in two scenarios: lack of large amount of training data, and out-of-domain translation (domain mismatch).

Unfortunately, large parallel corpora are only available for a limited set of domains such as news and political discourses. Several approaches for NMT domain adaption have been proposed [1], but these techniques require a small amount of parallel data or large quantity of monolingual data. This can be problematic for domains such as finance, where even monolingual data is scarce or non-existent due to the commercial value and privacy issue of such data.

The most commonly known and well studied English-French bilingual data is the train portion provided within the WMT'14 shared task (Bojar et al., 2014). It contains 40.8M sentence pairs extracted from five datasets that cover various domains: EUROPARL V7 (Koehn, 2005), UNITED NATIONS CORPUS (Eisele and Chen, 2010), COMMON CRAWL CORPUS, NEWS COMMENTARY, and 10^9 FRENCH-ENGLISH corpus.

Lison and Tiedemann (2016) present OPENSUBTITLES, a parallel corpus of 2.6 billion sentences across 60 languages originally mined from movies and Tv episode subtitles. The English-French portion of OPUS (Tiedemann, 2012) corpus contains a small collection (ECB) of parallel data in finance domain, that is collected from documents published by the European Central Bank.

PARACRAWL[2] is an ongoing project aiming to collect parallel data from the web for all 24 official European Union languages. The dataset is extremely large and noisy, therefore there have been endeavors to filter high quality pairs (Koehn et al., 2018). For example, even if the English-French subset contains over 4 billion sentence pairs, Ott et al. (2018) extracted 127M clean pairs after applying their filtering procedure. Some statistics about publicly available English-French corpora are given in Table 1.

Dataset	Domain	Sentences	Words
EUROPARL	politic	2.0	115.7
COMMON CRAWL	web	3.2	172.4
UNITED NATIONS	public	12.8	772.2
NEWS COMMENTARY	news	0.2	10.4
10^9 WORD	general	22.6	1479.6
OPENSUBTITLES	movies	32.6	521.0
OPUS-ECB	finance	0.1	12.2
SEDAR	finance	8.6	469.8

Table 1: Main characteristics of popular French-English publicly available parallel corpora, as well as SEDAR that we gathered in this work. Figures are in millions.

The main contribution of this work is the release of the SEDAR corpus, a large scale English-French parallel corpus for the financial domain. The corpus is assembled from public financial documents filed by Canadian issuers between 1997 and 2018. The parallel corpus will be made available for the research community by the *Autorité des marchés financiers du Québec* directly. Although it is limited to a single language pair, we hope that this endeavour will encourage other entities to share their publicly licensed data to the NLP scientific community, especially for domains were textual data is scarce.

We split SEDAR into train/valid/test sets, while ensuring the train/test overlapping ratio (see Section 3.6.) is at the same level as commonly used benchmarks. We run extensive experiments on NMT in order to study various aspects of our corpus and compare it with the general domain corpus of WMT'14 shared task (Bojar et al., 2014). We train NMT models on subsets of various sizes in order to measure:

Work done while the first author was interning at Autorité des marchés financiers (Québec).

[1] See (Chu and Wang, 2018) for a review

[2] `https://paracrawl.eu`

	Management Report of Fund Performance	News Releases	Financial Statements	Prospectus	Management's Discussion & Analysis
Document Pairs	69k (21%)	56k (18%)	54k (18%)	25k (8%)	15k (5%)
Avg. Word Count	4,389	1,273	9,478	21,152	10,532
Avg. Page Num	11	3	28	42	28
Avg. Table Num	12	1	20	35	24
Avg. Par. Num	168	32	463	580	321
% Short Par.	42%	25%	43%	26%	27%

Table 2: Statistics of the five most frequent document types in SEDAR. Short paragraphs are those that contains less than 7 words.

- the impact of train/test n-grams overlapping ratio on models performance.
- the quality of the in-domain (finance) and the out-of-domain (general-domain) translation.
- domain adaptation improvements from general-domain data to finance.

Our experimental results shows the importance of train/test overlapping ratio as an indicator to better understand generalization performance for NMT. Large scale in-domain data is crucial for NMT models' performance on financial domain, as systems trained on existing benchmarks perform poorly when tested on SEDAR. In addition, the results shows that selecting finance relevant sentences from general domain corpora can further boost the performance on SEDAR.

The remainder of the paper is organized as follows. In Section 2., we give an overview of the content of the original data, and the release. We describe the preprocessing and alignment process we applied at document, paragraph and sentence level in Section 3.. We report on experiments we conducted on neural machine translation in Section 4.. Section 5. discusses recent related works. We conclude in Section 6..

2. Data Collection and Release

The System for Electronic Document Analysis and Retrieval (SEDAR)[3] provides access to public security documents and information filed by issuers in Canada. The filings are made available for personal and non-commercial use only, and it is strictly forbidden1 to extract them with an automatic process (e.g. a crawler). The bulk of fillings are concentrated in recent years: documents submitted between 2014-2018 form 44% of the entire collection. Communications belong to 25 broad industrial groups, the five most frequent ones being financial services (17%), junior natural resource (12%), industrial products (9%), consumer products (8%), and metals and minerals (8%). Table 2 shows the main characteristics of the five most frequent document types in SEDAR. Statistics show that documents are long and contain a high portion of tables, which is challenging for any PDF converter toolkit (see Section 3.).

The data is the property of the Alberta Securities Commission on behalf of the Canadian Securities Administrators (CSA), the thirteen provincial and territorial Canadian securities regulatory authorities. The SEDAR corpus has been created in collaboration with the *Autorité des marchés financiers du Québec*. It is based on publicly-available documents and information filed in SEDAR between 1997 and 2018. The *Autorité des marchés financiers du Québec* will grant access to the SEDAR corpus without charge for academic research upon request[4].

3. SEDAR Creation

3.1. Document Alignment

The entire dump contains over 9 million PDF files, but only a small subset of 290k document pairs are indeed parallel. The reason behind this is that only issuers under the province of Quebec regulations are required to provide their documents in French.

In a nutshell, document alignment quality is very high. For most documents the language is provided with meta data. As of the rest of the documents, we group documents by issuers, perform language detection and align a pair of fillings if they have: different languages, the same type, they were emitted within two days, and have similar sizes and page numbers.

3.2. Text Extraction

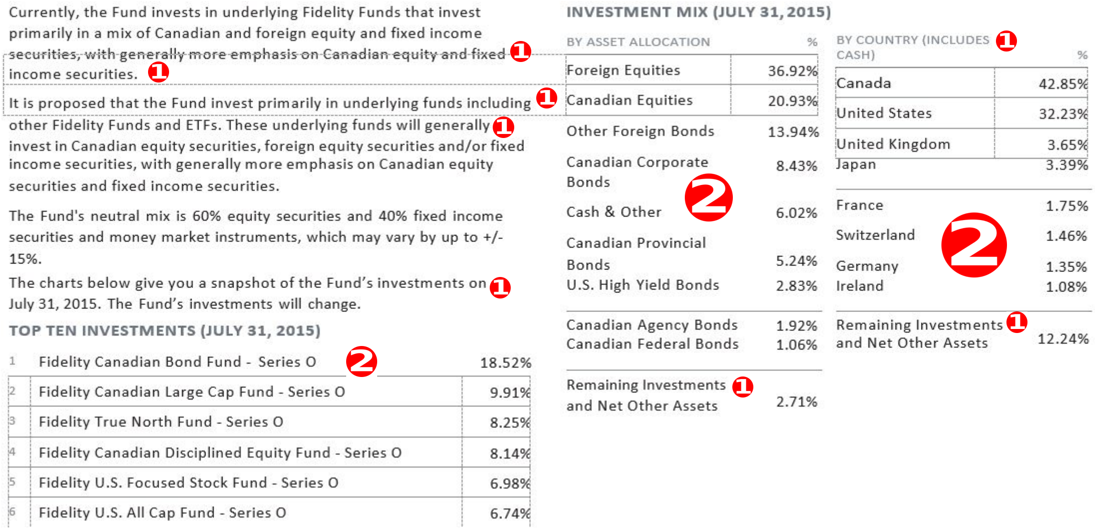
As PDFs were originally designed for individual investors, they are characterized by a complex layout, and a highly formatted and customized text structure and tables (e.g. tables without border). This makes text extraction an extremely hard task, where open source software like PDFBox (PDFBox, 2014) and Tika (Mattmann and Zitting, 2011) generate considerable amount of extraction errors and therefore cannot be reliably used for this type of documents.

Instead, texts were extracted using a commercial software (`Acrobat Pro DC`) which guarantees uniform, though not perfect, quality of the resulting MS WORD files. Converting 290k PDFs took 2 weeks using 5 modest computers (one node per machine).

Hence, the quality of extraction depends on the PDF converter, the filling year, as well as the complexity of the layout and formatting. Extraction errors includes: bad seg-

[3] https://sedar.com

[4] https://github.com/autorite/sedar-bitext



Currently, the Fund invests in underlying Fidelity Funds that invest primarily in a mix of Canadian and foreign equity and fixed income securities, with generally more emphasis on Canadian equity and fixed ❶ income securities. ❶

It is proposed that the Fund invest primarily in underlying funds including ❶ other Fidelity Funds and ETFs. These underlying funds will generally ❶ invest in Canadian equity securities, foreign equity securities and/or fixed income securities, with generally more emphasis on Canadian equity securities and fixed income securities.

The Fund's neutral mix is 60% equity securities and 40% fixed income securities and money market instruments, which may vary by up to +/- 15%.

The charts below give you a snapshot of the Fund's investments on ❶ July 31, 2015. The Fund's investments will change.

TOP TEN INVESTMENTS (JULY 31, 2015)

1	Fidelity Canadian Bond Fund - Series O ❷	18.52%
2	Fidelity Canadian Large Cap Fund - Series O	9.91%
3	Fidelity True North Fund - Series O	8.25%
4	Fidelity Canadian Disciplined Equity Fund - Series O	8.14%
5	Fidelity U.S. Focused Stock Fund - Series O	6.98%
6	Fidelity U.S. All Cap Fund - Series O	6.74%

INVESTMENT MIX (JULY 31, 2015)

BY ASSET ALLOCATION	%
Foreign Equities	36.92%
Canadian Equities	20.93%
Other Foreign Bonds	13.94%
Canadian Corporate Bonds	8.43%
Cash & Other ❷	6.02%
Canadian Provincial Bonds	5.24%
U.S. High Yield Bonds	2.83%
Canadian Agency Bonds	1.92%
Canadian Federal Bonds	1.06%
Remaining Investments ❶ and Net Other Assets	2.71%

BY COUNTRY (INCLUDES ❶ CASH)	%
Canada	42.85%
United States	32.23%
United Kingdom	3.65%
Japan	3.39%
France	1.75%
Switzerland ❷	1.46%
Germany	1.35%
Ireland	1.08%
Remaining Investments ❶ and Net Other Assets	12.24%

Figure 1: Excerpt from a PDF file with a complex layout converted to MS Word. Dashed lines show table borders recognized using the commercial toolkit we used. **Icon 1** indicates a badly placed line break, while **icon 2** represents unrecognized table structure. Parsing errors produce noisy text chunks that mislead publicly available sentence aligners.

ment boundaries; character decoding (especially French accents); gibberish text; reading order from text in multiple columns; layout and table recognition.

Figure 1 shows an excerpt of a converted PDF file, and parsing errors produced by the toolkit. The converted documents suffers from 2 main issues: truncated paragraphs (icon 1 in the figure) and unrecognized table structure (icon 2).

3.3. Alignment Methodology

As shown in Figure 2 (a), the raw document is structured in a sequence of blocks (paragraphs and tables), and the final goal is to produce parallel aligned sentences given a bilingual or pair of documents. In our case, the task is simplified by the fact that texts are already available in paragraph format.

Our first attempts to directly perform sentence splitting and alignment on the entire document generated a considerable amount of false positive alignments. It seems that the excessive presence of corrupted text chunks (discussed in previous section) biases popular sentence aligner toolkits.

Most sentence aligners (Gale and Church, 1993; Varga et al., 2007; Sennrich and Volk, 2010) are based on unsupervised algorithms making use of some sort of sentence similarity measures. Short text chunks that are partially similar constitute an important source of errors for those methods. After experimenting with multiple off-the-shelf sentence aligners, we realized we had to resort to a specific procedure.

By trial and error, we found satisfying to use a 2 pass option that first align paragraphs pairs using an iterative heuristic-based approach (Section 3.4.). In a second pass, we perform sentence splitting and alignment for each paragraph pair (Section 3.5.). Our approach is suited to texts characterized by a large number of tables and where the text are rather noisy and benefits from extra annotations provided in the MS Word format, such as font attributes.

3.4. Paragraph Alignment

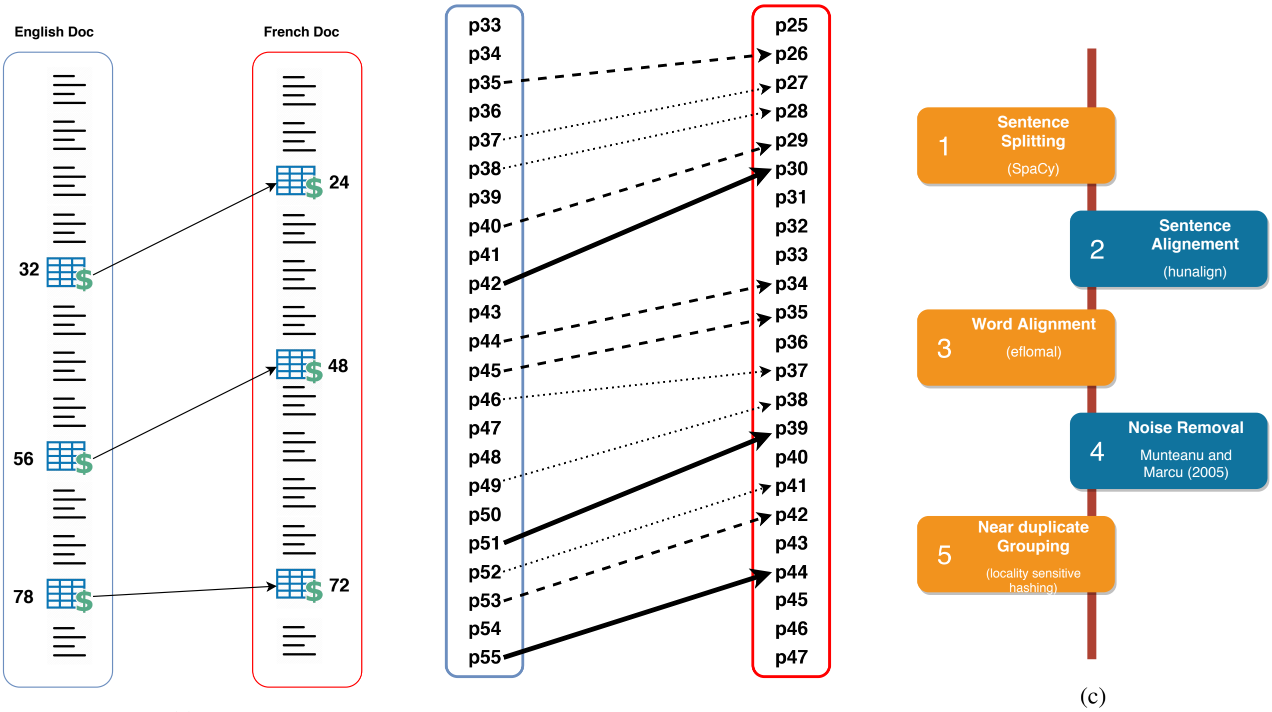
First, we align tables on both sides in order to build an initial set of potential paragraph pairs. As illustrated in Figure 2 (a), we use the position of the tables at each side as delimiters to limit possible candidate pairs. Tables can be aligned with a rather high accuracy using a simple heuristic that matches table dimensions and numerical values.

Instead of tackling a $n \times m$ alignment problem, we iteratively align subsets of potential paragraph pairs (Figure 2 (b)) from highest to lowest precision. At each iteration, we use the alignment as delimiters to update the set of potential candidate pairs. The main component of our iterative approach is a scoring function that scores a pair of paragraphs based on the percentage of words on each side that have a translation on the other side (according to a dictionary). In addition, the scoring function is powered by filtering rules that encourage:

1. font attributes (bold, italic, underline, size) matching
2. length (counted in words) matching (ratio of at most 2)
3. numbers, special characters, and punctuations matching

At early iterations, we use a high threshold with strict filters, which ensures high quality alignments (solid arrows in Figure 2 (b)) that consist of titles and clean – easy to align – paragraphs. Ordering the tiers from highest to lowest precision guarantees that most of the noisy candidate pairs are eliminated in early iterations.

At the end of each iteration, we use the index of the aligned pairs at each side to limit the set of potential candidates. As shown in Figure 2 (a), if the paragraphs pairs *42-30* and *51-39* are aligned, then English paragraphs *43-50* can only be aligned to *31-38* French ones. As shown in Figure 2 (b), no dashed arrows (alignment established after a few iterations) nor dotted ones (late iterations) can cross solid arrows.


English Doc
French Doc
32
56
78
24
48
72
p33
p34
p35
p36
p37
p38
p39
p40
p41
p42
p43
p44
p45
p46
p47
p48
p49
p50
p51
p52
p53
p54
p55
p25
p26
p27
p28
p29
p30
p31
p32
p33
p34
p35
p36
p37
p38
p39
p40
p41
p42
p43
p44
p45
p46
p47
1 Sentence Splitting (SpaCy)
2 Sentence Alignement (hunalign)
3 Word Alignment (eflomal)
4 Noise Removal Munteanu and Marcu (2005)
5 Near duplicate Grouping (locality sensitive hashing)

(a)

(b)

(c)

Figure 2: Processing workflow of our two pass approach for the creation of parallel corpus from the MS Word files. **(a)** Input Document is a sequence of paragraph and tables. We use aligned table pairs as delimiters to initialize a set of potential candidate paragraphs pairs. For example, English paragraphs [33,55] can only be aligned with [25,47] French ones. **(b)** Solid, dash and dotted arrows represent paragraphs alignments established at early, middle and late iterations respectively. At each iteration, we relax the alignment conditions, and update the set of potential candidates. **(c)** Sentence alignment procedure applied to each paragraph pair in order to produce the final parallel corpus.

At each iteration, we relax the rules by either reducing the scoring function threshold or by removing some filters. Relaxing the conditions will ensure a high recall, while controlling a decent precision. As we start with high confidence alignments, the number of noisy candidate pairs typically shrinks at each iteration, which reduces the number of non-parallel alignments at later iterations .

Surprisingly, despite the simplicity of our approach, manual inspection shows only a few number of non-parallel pairs, while most errors are correct aligned paragraphs with one of them being truncated. As most of the breaking lines are at the beginning or the end, these errors can be easily eliminated at the sentence alignment level as we shall see in the next section.

3.5. Sentence Alignment

Figure 2 (c) illustrates the processing steps used to produce alignment at sentence level. For each paragraph, we perform sentence splitting using `spaCy`[5]. Sentences are automatically aligned with `hunalign` (Varga et al., 2007). Then, extremely noisy sentence pairs (e.g. parallel but truncated sentences) are filtered.

We re-implemented the feature-based classifier proposed by Munteanu and Marcu (2005). The feature set includes: sentences lengths and ratio, largest three fertilities, longest connected and unconnected spans, number of words with no connection. `YASA` (Lamraoui and Langlais, 2013) was used to produce word alignments, which are required to calculate the value of many features used in the classifier. The word aligner and classifier are trained on high confidence alignments detected at early iteration of the paragraph alignment pass.

As the last step, we use locality sensitive hashing (Manku et al., 2007) in order to detect and group near duplicate sentences.

Preprocessed Corpus	**Number**
Total document pairs	0.29
Total paragraph pairs	47.0
Unique paragraph pairs	16.2
Total sentence pairs	70.9
Unique sentence pairs	17.2
after noise removal	12.0
after near duplicate grouping	8.6

Table 3: Sizes (in millions) of subcorpora generated at each preprocessing step.

Table 3 shows the evolution of the corpus in term of size at each preprocessing step. Paragraph alignment generates over 47M pairs, 65% of which being duplications. Manual inspection shows that repetitions are due to generic spans (e.g. *31 December*, *Short term bonds*), titles (e.g. *Credit*

[5] `https://github.com/explosion/spaCy`

Risk), specific application forms, and in some cases, issuers repeating an entire text block in their submissions. For example, it is common that news releases contain 1 or 2 pages of a company description that are repeated with each new submission. After sentence alignment and noise removal, we end up with 16.7M unique sentence pairs.

About 49% of these sentences can be retrieved and matched with at few number of substitutions as in the following sentence: *In 2008, net charges and adjustments increased the provisions by $6 million.* For each group of near duplicate sentences, we randomly pick one for the final corpus, which leaves us with roughly 8.6M unique and with no similar sentence pairs.

3.6. Train/Valid/Test Splitting

Because of the highly repetitive nature of the financial domain text style, a considerable number of sentences and expressions are repeated across years and issuers. Consequently, extracting entire documents as a test set would result in high overlap with the training material.

Arguably, a test corpus sampled uniformly over a one year period might be more representative. Therefore, we first reserve the 2018 fillings (450k sentence pairs) for validation and test. Then, we remove all sentences that overlap with previous years data.

We compute the percentage of n-grams[6] in the test set that already have been seen in the training set. We eliminate a test sentence if more than 10% of its 4-grams appear in the rest of the corpus. This way of splitting the corpus was deliberately chosen in order to recreate as much as possible the working environment of a system faced with the translation of unseen sentences. Also, splitting on this threshold makes train/test overlapping in our corpus consistent with commonly used benchmarks, this is discussed in more details in Section 4.2..

	`train`	`valid`	`test`
# sent pair	8.6M	6k	6k
# tokens	469.8M	264k	264k
# words	436k	27k	28k
# hapax	172k	10k	10k
word/sentence	27	23	21
Unseen words			
`sedar-train`	-	2.5%	1.7%
EuroParl	-	18%	18%
WMT'14	-	5%	5%

Table 4: Main characteristics of SEDAR subsets splitting: `sedar-train`, `sedar-valid`, and `sedar-test`. The bottom part shows the percentage of unknown words. For instance, 18% of token types of `sedar-test` are unseen in EuroParl.

Applying the filter reduces the number of no-overlapping sentences to 50k pairs, from which we create validation and test sets by randomly selecting 6k sentence pairs for each set. Table 4 shows the main characteristics of the three subset splits that we call from now on: `sedar-train`, `sedar-valid`, and `sedar-test`.

In order to measure the impact of train/test overlap, we generate another test set by randomly selecting 6k sentences from 2018 documents without overlap filtering. We call this subset `sedar-test-wof` and it is only used for contrasting results in Section 4.2..

4. Neural Machine Translation

4.1. Experimental setup

In addition to SEDAR subsets described in the previous section, we experiment with the English-French bilingual corpus provided by WMT'14 (Bojar et al., 2014) shared task, which is a widely used benchmark. We adopt `news-test-2012&2013` as a validation set, and `news-test-2014` as a test set.

In all experiments, we employ the Convolution model of (Gehring et al., 2017) as implemented in the `fairseq` toolkit (Ott et al., 2019) with the same configuration that the authors used for the WMT'14 English-French experiments.

Preprocessing of raw text is the same for all datasets used in this study, and it was carried out using the scripts accompanying the toolkit.

We evaluate the performance of our models using BLEU (Papineni et al., 2002), and report results on test of the best models performing on their corresponding validation set.

4.2. Impact of train/test Overlapping

In this experiment, we are interested in measuring the impact of train/test overlapping on NMT performance. Thus, we compare the performance of models trained on randomly picked 2M sentences from `sedar-train` and tested on 2 variants of the held-out datasets: before and after removing overlapping sentences. An additional system was trained on EuroParl (Koehn, 2005), and we report performance on `news-test`. Table 5 shows the performance of trained-models, and the overlap rate at 3-grams and 4-grams level between the 2 training sets (same size) and the three test sets: `news-test-2014`, `sedar-test-wof`, and `sedar-test`.

Expectedly, the performance measured on the overlapped `sedar-test-wof` is much higher than the one measured on the filtered version of the test set (`sedar-test`). Also, we observe a high overlapping rate between `sedar-train` and its test before filtering, compared with EuroParl and `news-test`. This is related to the nature of the financial domain text style, which involves a significant amount of near-duplicate sentences. This explains the BLEU score difference between the two configurations (30.65 on EuroParl vs. 51.79 on `sedar-train`). After filtering, the overlap ratio between `sedar-train` and its test drops dramatically to be close to the ratio between EuroParl and `news-test`.

Interestingly, the filtered `sedar-test` overlapped less with EuroParl. We observe a drop of 17% and 8% at the 3- and 4-gram levels respectively. This is because the filtered `sedar-test` consists of unseen sentences of 2018, while the last version of EuroParl was compiled in 2011.

[6] n-grams were generated on raw untokenized text.

Training Data	news-test-2014			sedar-test-wof			sedar-test		
	3-gm	4-gm	BLEU	3-gm	4-gm	BLEU	3-gm	4-gm	BLEU
EUROPARL (2M)	33%	12%	30.65	30%	10%	21.20	13%	2%	18.76
sedar-train (2M)	30%	9%	23.78	82%	68%	51.79	36%	7%	35.99

Table 5: Correlation between train/test overlap and the performance. NMT models are trained on EUROPARL and a subset of 2M randomly selected sentence pairs from `sedar-train`. The models are tested on: (a) `news-test-2014`, (b) `sedar-test-wof` (without overlap filtering), and (c) `sedar-test`. Performance is reported in BLEU columns, while 3-gm and 4-gm columns indicate the percentage of overlapping between training materials (rows) and at 3- and 4-grams levels respectively. The overlapping is calculated on the entire SEDAR corpus not on the 2M pairs.

Also, the filtering process removes the frequently used (domain independent) n-grams, which is reflected by a drop of BLEU from 21.20 to 18.76.

All those observations confirm the high correlation between train/test overlapping ratio and performances. This suggests that the train/test overlapping ratio is an interesting figure to report in case of experiments on new or multiple data sets.

4.3. Domain Adaptation

Our goal in this experiment is to measure how NMT models trained on general domain data (WMT'14) perform on financial data (SEDAR). First, we note that the overlapping ratio between WMT'14 and `sedar-test` is 34% and 8% at 3- and 4-gram level respectively, which is very close to that reported on SEDAR. This indicates that some material in WMT'14 is likely useful for handling financial domain texts.

We train NMT models on subsets of size $\in \{2, 4, 6, 8\}$ millions of randomly selected sentences pairs from `sedar-train` and WMT'14 (hereafter called WMT-RND). In addition, we experiment with *SEDAR domain adapted* subsets extracted from WMT'14 corpus, that we call WMT-SDA: we follow the approach of Axelrod et al. (2011) and Moore and Lewis (2010) to select sentences from a large general domain parallel corpus (WMT'14) that are the most relevant to the target domain (finance).

We train a 5-gram backoff language model using `kenLM` (Heafield et al., 2013) on `sedar-train` in order to score WMT'14 sentences. We generate finance domain-adapted training data by selecting the *k*-highest scoring sentences (lowest perplexity).

Left part of Figure 3 shows BLEU scores on `sedar-test` of NMT models trained on increasing subsets of `sedar-train`, WMT-RND, and WMT-SDA. Expectedly, models trained on `sedar-train` significantly outperform models trained on WMT'14. Also, models trained on WMT-SDA outperform randomly selected WMT-RND.

However, we observe that increasing training data size from 2 to 8 million reduces the gap between models trained on WMT-SDA and WMT-RND from 4 to 2 BLEU points. By inspecting LM scores, we noticed that at most the top 4 million pairs of WMT-SDA shares a high degree of similarity with SEDAR. The rest of sentences have similar low scores, which practically turns the selection process to a random selection.

In a last experiment, we investigate if general-domain data can further boost the performance of the best model on the financial domain (`sedar-test`). To this end, we concatenate `sedar-train` with increasing subsets of WMT-RND and WMT-SDA. As shown in the right side of Figure 3, we observe a boost in performance of 0.8 and 1.7 BLEU point when we add at most 8M sentence pairs from WMT-RND and WMT-SDA respectively. Expectedly, the gain with the latter is larger than with WMT-RND, confirming the validity of the selection process. That general domain data help on top of an abundant domain specific training set was not entirely expected, and is likely due to the variety of sources of texts available in the WMT'14 dataset.

5. Related Work

Due to privacy issues and the commercial value of financial domain data, publicly available datasets are too scarce and small to support NLP research in the field.

Lee et al. (2014) published a dataset of 13.7k (27.9M tokens) *8-K financial reports* extracted from the U.S. Security and Exchange Commission (SEC) filings and annotated with stock prices. The goal is to forecast companies stock price changes (*UP, DOWN, STAY*) using textual information of these financial reports. Interestingly, the authors demonstrate that textual data can insure gains of 10% to a strong baseline based on features crafted from numerical data.

In order to support adaption of Named Entity Recognition (NER) systems to the financial domain, Alvarado et al. (2015) annotated 8 *financial agreements* (54k tokens) with entity type labels: *LOCATION, ORGANISATION, PERSON, and MISCELLANEOUS*. Experiments conducted by the authors show that models trained on newswires (Tjong Kim Sang and De Meulder, 2003) perform poorly (17%) when tested on finance, compared to systems trained on in-domain data (83%).

Both studies suggest that large scale in-domain data are crucial in order to exploit the full strength of NLP techniques for financial domain related tasks.

In this work, we supply the scientific community with massive amount of parallel English-French data in the financial domain that can be directly used for machine translation. Furthermore, bitext data can also be used for word representation (McCann et al., 2017) and sentence embedding (Schwenk and Douze, 2017) learning.

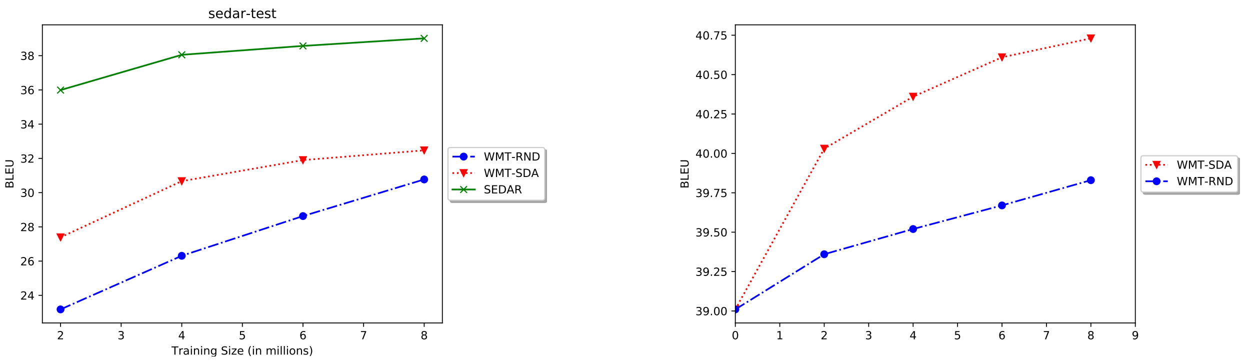


Figure 3: **Left** BLEU scores on `sedar-test` of models trained on increasing size subsets of: `sedar-train`, WMT-RND, WMT-RND corpora. **Right** Performances on `sedar-test` when `sedar-train` is incremented by subsets of WMT-RND and WMT-SDA. The x-axis represents the number of added sentences pairs (in millions). At point zero, only `sedar-train` (8M) is used for training.

6. Conclusion

We describe the acquisition and release of SEDAR, a large scale English-French bilingual corpus for the financial domain. Due to a high rate of false positives produced by standard sentence alignment techniques, we had to ressort to a dedicated strategy; suggesting that sentence alignment is not a solved problem.

We run experiments that shows the importance of train/test overlapping on machine translation systems evaluation. Also, we measure the impact of domain shifting on NMT performance, showing that large in-domain data is crucial to obtain good performances on finance. Furthermore, we improve the in-domain (finance) performance with a well known data selection process applied to the WMT'14 dataset.

While our corpus is restricted to a single language pair, we hope that this contribution will encourage governmental agencies and organizations to share publicly licensed data to the scientific community. We hope out resource will foster NLP research in the financial domain where datasets are currently very scarce.

7. Acknowledgements

The authors thank Lise Estelle Brault, François Mercier and Marylise Caron for their support for this project. The authors also thank Chantal Landry-Routhier for insightful discussion. This research was financed and supported by the Fintech Lab of the Autorité des marchés financiers (AMF), in order foster innovations for Québec financial sector. Furthermore, the authors thank all members of the Canadian Securities Administrators for the open approach for innovation and science. We thank the anonymous reviewers for their insightful comments.

8. Bibliographical References

Alvarado, J. C. S., Verspoor, K., and Baldwin, T. (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.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on empirical methods in natural language processing*, pages 355–362. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., et al. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chu, C. and Wang, R. (2018). A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Eisele, A. and Chen, Y. (2010). Multiun: A multilingual corpus from united nation documents. In *LREC*.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional Sequence to Sequence Learning. *ArXiv e-prints*, May.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Koehn, P., Khayrallah, H., Heafield, K., and Forcada, M. L. (2018). Findings of the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 739–752, Belgium, Brussels, October. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Lamraoui, F. and Langlais, P. (2013). Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment. *XIV Machine Translation Summit*.

Lee, H., Surdeanu, M., MacCartney, B., and Jurafsky, D. (2014). On the importance of text analysis for stock price prediction. In *LREC*, pages 1170–1175.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Manku, G. S., Jain, A., and Das Sarma, A. (2007). Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150. ACM.

Mattmann, C. and Zitting, J. (2011). *Tika in action*. Manning Publications Co.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. (2017). Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6297–6308.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*, pages 220–224. Association for Computational Linguistics.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

PDFBox, J. (2014). processing library. *Link: http://www. pdfbox. org*.

Schwenk, H. and Douze, M. (2017). Learning joint multilingual sentence representations with neural machine translation. *arXiv preprint arXiv:1704.04154*.

Sennrich, R. and Volk, M. (2010). Mt-based sentence alignment for ocr-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.