Universität Zürich[UZH]

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Arts
der Philosophischen Fakultät der Universität Zürich

A Task-specific Neural Translation Model for the Stellenmonitor Pipeline

Verfasser: Silvio Daniele Magaldi
Matrikel-Nr: 14-921-936

Referent: Prof. Dr. Rico Sennrich
Betreuerin: Chantal Amrhein
Institut für Computerlinguistik

Abgabedatum: 01.12.2021

Abstract

The aim of this work is to provide a multilingual task-specific machine translation model to the pre-existing text zoning and classifier pipeline by Gnehm and Clematide (2020) for job advertisements of the Swiss Job Market Monitor. The main challenge of such a model is the very low amount of in-domain parallel data available for the task. Thus, this work is investigating the efficacy of different domain adaptation methods in a low resource domain on three language pairs: English > German, French > German, and Italian > German. Due to time constraints, in a first step, these methods are limited to (1) raw fine-tuning using in-domain data only, (2) concatenating consecutive segments, (3) supporting fine-tuning with out-of-domain data, and (4) Elastic Weight Consolidation. The performance of the different methods will be measured mainly by the classifier's performance on their translation. BLEU is considered as a supporting measure. In a second step, the best previous approach will be used to produce and then evaluate the efficacy of (5) back-translation and (6) forward-translation as synthetic parallel data for domain adaptation. It can be shown that domain adaptation supported with generic out-of-domain data is the best performing method for this low in-domain resource task. Furthermore, it can be shown that the use of back-translations is beneficial for this task. Whilst the addition of forward-translations boosts the BLEU-score of translations of all three languages, when taking all three languages into account, the classifier performance is lowered.

Zusammenfassung

Das Ziel dieser Arbeit ist es, ein mehrsprachiges, aufgabenspezifisches maschinelles Übersetzungsmodell für die bereits existierende Text-Zoning- und Klassifizierungs-Pipeline von Gnehm und Clematide (2020) für Stellenanzeigen des Stellenmarkt-Monitor Schweiz bereitzustellen. Die sehr geringe Menge an parallelen Daten, welche für diese Aufgabe zur Verfügung stehen, ist die größte Herausforderung dieses Modells. Daher wird in dieser Arbeit die Wirksamkeit verschiedener Domänenanpassungsmethoden in einer Domäne mit geringen Ressourcen an drei Sprachpaaren untersucht: Englisch > Deutsch, Französisch > Deutsch, und Italienisch > Deutsch. Aus Zeitgründen beschränken sich diese Methoden in einem ersten Schritt auf (1) rohes Fine-Tuning unter ausschließlicher Verwendung von In-Domain-Daten, (2) Verkettung aufeinander folgender Segmente, (3) Fine-Tuning unterstützt mit Out-of-Domain-Daten und (4) Elastic Weight Consolidation. Die Leistung der verschiedenen Methoden wird hauptsächlich anhand der Leistung des Klassifikators auf deren Übersetzung gemessen. BLEU wird als ein unterstützendes Maß verwendet. In einem zweiten Schritt wird der beste vorherige Ansatz verwendet, um synthetische Paralleldaten zu produzieren und um die Wirksamkeit der Verwendung von (5) Back-Translation und (6) Forward-Translation für die Domänenanpassung zu untersuchen. Es kann gezeigt werden, dass die Domänenanpassung, unterstützt durch generische Out-of-Domain-Daten, die beste Methode für diese Aufgabe mit wenig In-Domain-Ressourcen ist. Außerdem kann gezeigt werden, dass die Verwendung von Back-Translation für diese Aufgabe von Vorteil ist. Obwohl die Hinzunahme von Vorwärtsübersetzungen den BLEU-Score der Übersetzungen aller drei Sprachen erhöht, sinkt die Leistung des Klassifikators, wenn alle drei Sprachen berücksichtigt werden.

Acknowledgement

I want to thank Ann-Sophie Gnehm for providing the in-domain data and the Swiss Job Market Monitor's classifier for this work in a well structured manner, taking the time to provide an introduction to this data and the classifier, and being available for questions throughout this work. Furthermore, I want to thank the Department of Informatics, as well as the Department for Computational Linguistics, at the University of Zurich for providing access to one of their GPU-servers.

Contents

Abstract i

Acknowledgement ii

Contents iii

List of Figures vi

List of Tables vii

1 Introduction 1
- 1.1 Problem Statement . . . 1
- 1.2 Research Objectives . . . 1

2 Research Context 2

3 Approach 4
- 3.1 Data . . . 4
  - 3.1.1 Generic Data . . . 4
  - 3.1.2 In-domain Data . . . 5
- 3.2 Data Preparation . . . 6
  - 3.2.1 Human Translations . . . 6
  - 3.2.2 Monolingual In-domain Data . . . 7
- 3.3 Evaluation . . . 7
- 3.4 Framework . . . 8
- 3.5 Task Overview . . . 8

4 Baseline Results Covering Human Translation Set: Human Translation vs. DeepL-translation 10
- 4.1 Raw Baseline . . . 10
  - 4.1.1 French . . . 10
  - 4.1.2 English . . . 10
- 4.2 Analogous Pre-Processing on DeepL-translation . . . 12
- 4.3 Unprocessed Human Translation . . . 13
  - 4.3.1 French . . . 14
  - 4.3.2 English . . . 14
- 4.4 Best Baseline . . . 15
  - 4.4.1 Class Composition of the Gold Standard . . . 15
  - 4.4.2 DeepL Baseline for Final NMT Models . . . 15

5 Model Variants 18
- 5.1 Generic Base Model . . . 18

5.2 Alpha Model Variants . . . 18
5.2.1 Alpha Variant 1: Fine-Tuning on Single Segments . . . 18
5.2.2 Alpha Variant 2: Fine-Tuning on Two Concatenated Segments . . . 19
5.2.3 Alpha Variant 3: Fine-Tuning Supported by Generic Data . . . 19
5.2.4 Alpha Variant 4: Fine-Tuning Using Elastic Weight Consolidation . . . 20

6 Alpha Model Results 21
6.1 Generic Base Model . . . 21
6.1.1 English . . . 21
6.1.2 French . . . 21
6.1.3 Italian . . . 21
6.2 Alpha Variant 1: Fine-Tuning on Single Segments . . . 21
6.2.1 English . . . 22
6.2.2 French . . . 22
6.2.3 Italian . . . 22
6.3 Alpha Variant 2: Fine-Tuning on Two Concatenated Segments . . . 22
6.3.1 English . . . 22
6.3.2 French . . . 23
6.3.3 Italian . . . 23
6.4 Alpha Variant 3: Fine-Tuning Supported by Generic Data . . . 23
6.4.1 English . . . 23
6.4.2 French . . . 23
6.4.3 Italian . . . 24
6.5 Alpha Variant 4: Fine-Tuning Using Elastic Weight Consolidation . . . 24
6.5.1 English . . . 24
6.5.2 French . . . 24
6.5.3 Italian . . . 25
6.6 Alpha Variants Conclusion . . . 25
6.7 Constrained Decoding . . . 26

7 Synthetic Data Produced 27
7.1 Back-translation from German . . . 27
7.2 Forward-Translation to German . . . 27

8 Models Using Synthetic Data 28
8.1 Variant 1: Using All Back-translations . . . 28
8.1.1 Variant 1.1: Supported by Half of the Generic Data. . . . 28
8.1.1.1 English . . . 28
8.1.1.2 French . . . 28
8.1.1.3 Italian . . . 29
8.1.1.4 Extended Domain-Adaptation . . . 29
8.1.2 Variant 1.2: Supported by a Quarter of the Generic Data. . . . 29
8.1.2.1 English . . . 29
8.1.2.2 French . . . 30
8.1.3 Italian . . . 30
8.2 Adding Forward-translations . . . 30
8.2.1 Variant 2.1: Forward-translations Added to Variant 1.1 . . . 30
8.2.1.1 English . . . 30

8.2.1.2 French . . . 30
8.2.1.3 Italian . . . 30
8.2.2 Variant 2.2: Forward-translations Added to Variant 1.2 . . . 31
8.2.2.1 English . . . 31
8.2.2.2 French . . . 31
8.2.2.3 Italian . . . 31
8.3 Variant 3: Elastic Weight Consolidation . . . 31
8.4 Comparing Variants across Languages . . . 32
8.4.1 Result Validation on the Test Set . . . 33
8.4.1.1 Class Distribution on the Test Set . . . 33
8.4.1.2 Results on the Test Set . . . 34

9 Conclusion **36**
9.1 BLEU as a Performance Predictor . . . 36
9.2 Limitations . . . 36
9.3 Next Steps . . . 37

List of Figures

4.1 F_1-scores on DeepL-translations pre-processed English > German 12
4.2 F_1-scores on DeepL-translations pre-processed French > German 12
4.3 F_1-scores on DeepL-translations and Human Translations, with and without duplicates. English > German 14
4.4 F_1-scores on DeepL-translations and Human Translations, with and without duplicates. French > German 14
6.1 Unweighted F_1-scores on Alpha Variants English > German 25
6.2 Unweighted F_1-scores on Alpha Variants Italian > German 25
6.3 Unweighted F_1-scores on Alpha Variants French > German 25
8.1 Unweighted F_1-scores on Variants using synthetic parallel data on the development set English > German 32
8.2 Unweighted F_1-scores on Variants using synthetic parallel data on the development set French > German 32
8.3 Unweighted F_1-scores on Variants using synthetic parallel data on the development set Italian > German 33
8.4 Unweighted F_1-scores on Variants using synthetic parallel data on the test set English > German 35
8.5 Unweighted F_1-scores on Variants using synthetic parallel data on the test set Italian > German 35
8.6 Unweighted F_1-scores on Variants using synthetic parallel data on the test set French > German 35

List of Tables

3.1 Number of parallel segments per source language; per corpus and total. 4
4.1 Classification results on raw human translations and DeepL-translations from French to German. Italic: scores below 0.25 11
4.2 Classification results on raw human translations and DeepL-translations from English to German. Italic: scores below 0.25, bold: greatest difference 11
4.3 Classifier results on fully processed and lower-cased DeepL-translation, English>German 13
4.4 Best classifier results of the DeepL-translation to be used as a benchmark and class distribution on the gold standard in percent, English > German 16
4.5 Best classifier results of the DeepL-translation to be used as a benchmark and class distribution on the gold standard in percent, French > German 17
4.6 Best classifier results of the DeepL-translation to be used as a benchmark and class distribution on the gold standard in percent, Italian > German 17
6.1 Weighted average F_1-score compared between constrained and unconstrained decoding. Based on Alpha Variant 3: Fine-Tuning Supported by Generic Data. (See 5.2.3) 26
8.1 Distribution of job classes in gold-standard data across all Sets in percent 33

1 Introduction

1.1 Problem Statement

The aim of this work is to provide a task-specific machine translation model to the pre-existing text zoning and classifier pipeline by Gnehm and Clematide (2020) for job advertisements of the Swiss Job Market Monitor (hereafter SJMM). Currently, most labeled data to train this pipeline is only available in German. As Switzerland is a multilingual country though, there are many advertisements in French and Italian. In more recent years, more and more advertisements are also published in English, especially online. Thus, Gnehm and Clematide decided to experiment with different approaches. One of these approaches was handling these other languages by using a multilingual classifier. Another one was using the generic online machine translation tool by DeepL. These approaches were, however, not entirely satisfactory. Thus, this work will aim to create a domain-specific machine translation model which should outperform generic machine translation models such as DeepL based on classification scores of the previously mentioned pipeline.
The main challenge of this work is the small amount of in-domain training data. There are few human-translated job advertisements in English and French, and none for Italian.

1.2 Research Objectives

Based on the previous section, my aim in this work is to train such a domain- and task-specific translation model. As mentioned, there is very little to no in-domain training data. Thus, I will use a multilingual machine translation model to translate bidirectionally between English and German, French and German, and Italian and German. This should, on the one hand, allow for domain adaptions in the directions English > German and French > German to be transferred to Italian > German as well. On the other hand, this will enable the use of back-translation and forward-translation in order to create new synthetic in-domain training data (Sennrich et al., 2016a). This should allow for improved performance of SJMM's existing pipeline on originally non-German input, enabling the better use of the collected data.
The performance of my translation models will mainly be measured by the performance of this classification pipeline as this is its intended use and allows for comparison with the generic machine translation model of DeepL.
This work is only concerned with providing better translated German input to SJMM's pipeline; I will in no way alter said pipeline.

2 Research Context

As only very few human-translated in-domain segments are available, the first focus should lie on creating synthetic in-domain parallel training data. As previously mentioned, it makes sense to use back-translation for this purpose (Sennrich et al., 2016a; Edunov et al., 2018). In this case, this means translating German advertisements to English, French, and Italian. The resulting synthetic parallel data has very high quality on the German side. Additionally, it could be beneficial to also use a small part of forward-translation, as it was shown by Bogoychev and Sennrich (2019) that whilst back-translation is the more beneficial approach, a combination of back- and forward-translation can improve translation performance. Human evaluators mainly criticised the lower fluency in forward-translated segments compared to back-translation (Bogoychev and Sennrich, 2019). But since this translation model will work on job advertisements where fluency tends to be lower with more fragmented segments, this may not be an issue at all.
For this work, I use Byte Pair Encoding, short BPE (Sennrich et al., 2016b). By using this method, a shared infinite vocabulary can be built for all four languages covered by this work as each word can be built of its subwords. This is especially important as some key words in job advertisements, such as the job title itself, may be rare words. Without the use of subwords, these rare but highly important words would, especially in a multilingual vocabulary, result in unknown tokens.
In this task, the system's use is strictly limited to job advertisements. Therefore, domain adaptation makes sense. It has to be kept in mind though, that most of the in-domain data will contain fragmented sentences due to the nature of job advertisements. Such sentence fragments may be problematic, as, taken on their own, they lack crucial context for the translation. For this work, I thus will train the system to translate based on the concatenation of consecutive segments for context extension. I assume that this will especially alleviate the issue of lacking context surrounding associated sentence fragments. Based on the previous work by Lopes et al. (2020), this approach looks promising concerning general performance as well. In the previously mentioned paper, the "2to2" concatenation approach looks the most promising. In this "2to2" approach, two consecutive segments are concatenated and the segment boarded is indicated by a special token. This concatenation is performed for the source and for the target side as well.
As the domain adaption is performed on the basis of a generic machine translation model, catastrophic forgetting has to be kept in mind. Even though the proposed machine translation model is not intended to ever handle other translation tasks, in a first step, a machine translation model to create back-translations and forward-translations must be trained. There, the very small amount of natural in-domain data available would most likely lead to overfitting very quickly. The need for limiting the impact of domain adaptation is further supported by the fact that, for the training of the final translation model, most available in-domain data will be synthetic parallel data (back-translated and forward-translated). For this purpose, I will investigate the efficacy of (1) supporting the in-domain data with generic data but oversampling the in-domain data. This was shown by Chu et al. (2017) to be a highly promising approach. I will then compare this method to (2) Elastic Weight Consolidation during domain adaptation. In this approach, the empirical Fisher information is calculated for the translation model's

weights. This then allows to avoid changing weights that are crucial for the performance on generic translations during domain adaption (Thompson et al., 2019).

In the data that was translated with DeepL, it can be seen that there are terminology errors. These are especially prevalent with abbreviations concerning education levels. Some of these terms are specific to Switzerland and some are outdated. As previously mentioned, this task lacks high-quality in-domain parallel data and thus, this terminology cannot be acquired through domain adaptation at all. Additionally, outdated terms are unlikely to appear in any of the corpora. Constraint based decoding (Post and Vilar, 2018) is a reasonable choice to investigate.

3 Approach

3.1 Data

3.1.1 Generic Data

In a first step, the generic translation model will be trained on three corpora. The first corpus is the Europarl corpus containing transcripts and translations from the European Parliament, thus providing varied domain parallel training data, generally of high quality (Koehn (2005); Tiedemann (2012)). For the two language pairs German and English, as well as German and French, I used the Europarl data released for the WMT21 shared task. This was done in hopes that the amendments performed by the organizers would also be beneficial to this task. For the language pair German and Italian, I used the OPUS release of said corpus. As a second corpus, I used the Tatoeba corpus (Tiedemann (2012)). This corpus should provide a solid basis of, though short, high-quality training data. Finally, I used the ECB corpus containing data from the website and documentation of the European Central Bank (Tiedemann (2012)). This decision was made as 11% of job advertisements in the SJMM's data that are labeled are from the financial sector. Thus, it should be useful for the translation model to be provided bank-specific data as such terms could also be key words in SJMM's classification pipeline.
Should the translation model have trouble with the sentence fragments present in job advertisements, I consider the TED2020 corpus as additional data (Tiedemann (2012)). This corpus contains transcripts and translations of TED talks, meaning spoken language which is also more fragmented than written language.
Whilst the scope of this work does not allow for it, it would be beneficial to also examine potential benefits from adding the News-Commentary corpus (Reimers and Gurevych (2020)) and version 8 of the ParaCrawl corpus (Bañón et al. (2020); Tiedemann (2012)). According to the maintainers of the ParaCrawl corpus, their focus for version 8 lay on its cleaning which makes it more interesting for generic model training. In a continuation of this work, increasing the amount of data for the generic machine translation model may be one source of further improvement.

Number of Parallel Segments per Source Language			
Corpus	English	French	Italian
ECB	100'000	100'000	100'000
Europarl	1'803'347	1'803'347	1'800'264
Tatoeba	307'333	110'548	22'292
Total	2'210'680	2'013'895	1'922'556

Table 3.1: Number of parallel segments per source language; per corpus and total.

3.1.2 In-domain Data

Regarding in-domain data provided by SJMM, there are 361 unique monolingual human-translated job advertisements. Of these 361 advertisements, 332 are English > German, and 29 are French > German. This results in 10083 segments for English > German, and 350 segments for French > German.
Additionally, there are 31 advertisements tagged as unknown. These are multi-lingual advertisements. The multi-linguality does, however, differ vastly: In some cases, it is limited to simple key dates such as place of employment etc.; in other cases, the company profile is in a different language than the ad; and finally, some advertisements are indeed bilingually written. Due to the scope of this work, these advertisements were not used. In a continuation of this work, further improvement may be achieved by sorting and cleaning these advertisements. The truly bilingual advertisements may be sources of parallel data. For this, each would have to be examined in order to make sure that the two texts are indeed translations of each other and not just two texts with equivalent content.
These human translations were only available as pre-processed texts. This pre-processing was performed due to a limitation in a, now obsolete, module within the SJMM's pipeline. One step of this pre-processing was the lower-casing of the texts. This may be especially detrimental in German, where nouns are capitalised. A general detriment of the loss of uppercase is the lack of discernability of acronyms which may cause unwanted side effects in translations. Additionally, Umlauts were substituted with their digraphic representation, e.g. «ä» is substituted with «ae». Whilst to a (native) speaker of German this is no problem at all, for the classifier and for the translation system this may cause issues. Another issue is posed through an encoding error: In some advertisements, the remaining diacritic symbols and the Unicode EN DASH (U+2013): – contained in the texts are replaced with the actual Unicode replacement character (U+FFFD): . This impacted mainly French advertisements. E-mail addresses and URLs were masked with a special token; this should be beneficial for performance (Gnehm and Clematide, 2020).
In addition to these human-translated advertisements, there are also a lot of advertisements translated using DeepL with gold standard tags provided by SJMM. Due to DeepL's terms and conditions[1], these translations may, however, not be used in order to train a translation model. I will use this data as a baseline comparison.
Most of the SJMM's data is in the form of monolingual job advertisements. There is a great amount of data in all four languages covered in this work though. These monolingual advertisements can be used to create synthetic parallel training data. There are unique gold standard labels for 3278 English, 5338 French, and 708 Italian advertisements. Of these 400 French, 350 English, and 74 Italian advertisements will be reserved as the development and the test sets respectively for the evaluation of the new neural machine translation models. Of the 74 randomly selected advertisements of the Italian development set, the DeepL API failed to translate two of the advertisements.
As a domain-specific terminology, I will be using the Swiss Standard Classification of Occupations CH-ISCO-19 (BFS, 2021a) provided by the Federal Statistical Office in German, French, and Italian. For training, the less detailed version thereof resulted in 485 phrase and single word translations for every language here considered (de, fr, en, it). To use these terms and phrases as a glossary, to be used during constraint based decoding, the selection was much more conservative in order to avoid false terminology enforcement. The removed entries were mainly

[1] https://www.deepl.com/en/pro-license#pro

multi-word phrases. Additionally, there is an extensively detailed version of said publication (BFS, 2021b) including gender variants. These are 34'763 job titles that occur in German, French and Italian. Male and female variants are counted as two job titles. Due to the size of the detailed variant and the lesser gender differences in English, I decided against translating these into English as well due to the expected low cost benefit ratio. In order to improve the performance on abbreviations of education levels, I will utilise abbreviations of certificates and degrees used in Switzerland in the same three languages. In this case, I again decided against translating these abbreviations to English. The reason being, that in the inspected advertisements the abbreviations in question are either translated to a descriptive phrase for equivalent qualifications and in some cases, the Swiss abbreviation is added in the local language of the advertising business.

3.2 Data Preparation

3.2.1 Human Translations

As described in section 3.1.2, the human translations contained in text files were significantly modified. In order to get better training data, especially in German where nouns are capitalised, this needed to be reverted.
First, the digraphic Umlauts were substituted with the diacritic Umlauts automatically using a custom but simple script. This script first substituted all Umlaut candidates «ae», «oe», and «ue». As e.g. not every «ae» is meant to be an «ä», in a second step, the exceptions were reverted back to «ae» by mapping the false substitution onto the correct (sub-) word using a dictionary. Examples of such exceptions are «Michael», «eventuell». I first developed this dictionary theoretically and refined it empirically by checking the results. As this is a small data set, I do not claim this script would work for different data sets.
As previously mentioned, the Unicode replacement characters were all the same characters on the byte level, thus the easiest, and due to the low amount of data and the low number of occurrences the fastest, way was to search them automatically and replace them by hand.
The human translations having been lower-cased meant that the text needed to be truecased. Additionally, the full human translation was a single line, whilst the original advertisements and the DeepL-translations were kept in the original segmentation as collected by the crawler. In order to create parallel training data, I decided to break down the single line back to the original segmentation. For this purpose, I tried different methods with different automatic alignment tools. The results were, however, entirely dissatisfactory as much of the data was discarded. Finally, I chose to use Explosion's `spaCy` for both of these issues (Honnibal and Montani, 2017). In order to truecase the text, I used `spaCy`'s part of speech tagger and capitalised proper nouns, and in the case of German also capitalised nouns. The tagger missed only a few nouns in German when multiple nouns followed each other directly. In order to re-segment the text, I used the senter of the same `spaCy` pipeline to predict sentence boundaries. Upon inspection, the results looked promising. The segmentation was consistent and different languages provided highly comparable results. Thus, this approach seems viable.

3.2.2 Monolingual In-domain Data

The raw crawled monolingual in-domain data is available in XML format where each crawled line is its own entry. This includes many empty lines that need to be ignored. Additionally, whilst most entries are a single segment of a job advertisement, there are entries that are composed of multiple segments. In some cases, job advertisements are even comprised of only one entry but this entry contains a multitude of segments. Faced with these variants, I have decided that the most consistent approach will be to concatenate the full advertisement into one line. Analogous to the human translation, I then let `spaCy`'s senter predict sentence boundaries, and according to these predicted boundaries, the texts were re-segmented. This way, there is no large discrepancy between the training data and the data seen during the inference task.

In terms of cleaning, I used the same e-mail address and URL masking as was applied to the human translations (Gnehm and Clematide, 2020). In some cases, this led to multiple consecutive masking tokens. In most of these occurrences, it was obvious that it should have been one:

(3.1) original: https://www.gowest-aarau.ch/

(3.2) processed: [URL][DNSHost]

(3.3) target state: [URL]

These masking tokens were reduced to a single one using a custom cleaning script. This cleaning script also removed a diverse array of undesired elements. These undesired elements were e.g. HTML elements including their tags which the crawler collected:

(3.4) raw: </div> <div id="profil">Profil minimum requis :&[...]

(3.5) processed: Profil minimum requis : [...]

After these processing and cleaning steps, BPE was applied using subword-nmt[2] with 32'000 merges, resulting in a vocabulary size of 32'795. For multi-segment experiments, I concatenated consecutive segments after applying BPE. These segments are separated by the special token `<seg>`:

(3.6) raw (1): Duration: 1,5 years, possibility of extension

(3.7) raw (2): More information at: [wwwadr] or [wwwadr] or Telefon [tel] (ITECO) or [tel] (IC)

(3.8) processed : Duration : 1,5 years , possibility of extension <seg> More information at : [wwwadr] or [wwwadr] or Telefon [tel] (ITECO) or [tel] (IC)

3.3 Evaluation

As this project aims at improving classification results on the pipeline of SJMM, the main measurement will be the classification score. BLEU-scores will be considered as a supporting measurement. Whilst its significance drops with regards to high-performance MT-Systems, in my personal experience it still gives an indication of improvement or deterioration with regards

[2] https://github.com/rsennrich/subword-nmt

to similar models. Additionally, better translated output may allow for improvement through retraining of pipeline components and thus, the linguistic quality of the output should also be considered to some extent. Potential retraining may be necessary due to a different distribution regarding classification labels of English job advertisements (Gnehm and Clematide, 2020). In order to calculate the BLEU-score, I use sacrebleu's[3] corpus_bleu function (Post, 2018). One problem with calculating a BLEU-score is that, as previously mentioned in section 3.1, this task lacks parallel data. The little parallel data I have in the form of the human translations, I use completely in the training of the models. Thus, I cannot use them as evaluation data. However, there still is the DeepL translated data; whilst I must not use it as training data, I may use it to calculate a BLEU-score against it. As previously mentioned, the absolute value of the BLEU-scores will not be used to make any decision. However, since DeepL-translations mostly excel at fluency, I will mainly interpret the difference in BLEU-score as a representation of fluency and readability to humans. I will also observe the BLEU-score in relation to the classifier performance.
SJMM's classifier classifies job advertisements into one of eleven job fields. SJMM's evaluator then checks the classifier's predictions against the gold standard. It creates an evaluation report using the python `sklearn.metrics.classification_report`[4] function. In a first part, it indicates precision, recall, F_1-score, and the number of advertisements used to support these numbers per class. In a second part, it indicates the overall accuracy, the macro average, the weighted average, and again the number of advertisements used to support these numbers. "The reported averages include macro average (averaging the unweighted mean per label), weighted average (averaging the support-weighted mean per label), [...]." (Pedregosa et al. (2011))

3.4 Framework

As I aim to use elastic weight consolidation, I use awslabs' Sockeye 2.0[5] (Hieber et al., 2020) as a framework which already has a draft implementation thereof. (Model details in section 5. For hyperparameters see Appendix: Hyper Parameters.)

3.5 Task Overview

The first step in this work is to create a baseline evaluation. For this purpose, I run the job field classifier on the human-translated data and on the DeepL translated data. As described in section 3.1.2 and section 3.2.1, the human translations were subject to different pre-processing steps. Thus, in this step, I also examine the effect of these different pre-processing steps on the classifier performance when applied to the DeepL-translations. In order to reduce computation time, I will base different domain adaptation approaches on a pre-trained generic bidirectional multilingual translation model. These different domain adaptation approaches will make use of the human-translated data available (see section 3.1.2 In-domain Data) and thus create a first set of domain-specific translation models. I will call these models the alpha models. There are three language pairs within the model: English and German, French and German, and finally

[3]https://github.com/mjpost/sacrebleu
[4]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html
[5]https://github.com/awslabs/sockeye

Italian and German. These language pairs are trained bidirectionally. The reason for this design decision is that I will use the best performing alpha model on the development set to create back-translations of the unlabeled monolingual data. These back-translations, together with the human translations, will then be used to train the second set of domain-adapted translation models from the pre-trained generic model. I expect a noticeable performance increase due to this massive increase in synthetic in-domain training data. For this second set, I will only use the top two approaches from the alpha set, in order to limit the scope of this work. Additionally to using back-translations, I will also train a set of models which include a smaller amount of forward-translations. Finally, I will compare these models' performance with the classifier on the test set.

The concrete steps:

- Perform a baseline evaluation of human-translated advertisements and DeepL translated advertisements.
- Get data ready for usage in training and translation.
- Train a baseline translation model.
- Adapt baseline model using human-translated advertisements (different methods: only target domain, target domain supported by generic data, only target domain using Elastic Weight Consolidation).
- Evaluate adapted models and select best on dev set classifier performance.
- Back-translate advertisements to be used as synthetic parallel data.
- Adapt baseline model using human-translated and back-translated advertisements.
- Evaluate different domain adaptation approaches (best method(s) found in previous steps).
- Using constrained decoding for translations.
- Select best model on development set for back- and forward-translations.
- Perform back- and forward-translations.
- Train final model variants and compare performance on test set.

4 Baseline Results Covering Human Translation Set: Human Translation vs. DeepL-translation

4.1 Raw Baseline

In order to have a baseline to compare the performance of my translations to, I first ran the SJMM classifier on the raw human translations as provided by the SJMM (Gnehm and Clematide, 2020) and on the original DeepL-translations with one alteration; since the human translation was tokenised, I also tokenised the DeepL data using spaCy (Honnibal and Montani, 2017). Afterwards, I examined the effects of different pre-processing steps on the DeepL-translation set. And finally, I ran the classifier on the re-processed human translation. As mentioned in section 3.1.2, these raw data contain 30 originally French advertisements and 368 originally English advertisements. There are no advertisements translated from Italian. Of the eleven available labels, only eight are present in the French data set. In the English data set, ten of eleven labels are present. The label which does not occur in the gold standard for both sets of advertisements is "Construction". In the French data set, "IT", and "Management & Organisation" do not occur.

4.1.1 French

In the French dataset, the raw human translation has enabled slightly better overall classifier performance than the DeepL-translation (see table 4.1). On two labels, the classifier performed better on the DeepL-translation, these being "Hospitality & Personal Services" where the recall was better (difference of one advertisement only) and "Teaching & Public Services" where the precision was better at the cost of recall. Due to the low number of advertisements and these two labels being more common, this gives the DeepL-translation the better weighted average.

4.1.2 English

For the English data set with more than ten times the amount of data, the performance difference is much clearer. Here, the DeepL-translation enables much better performance than the raw human translation (see table 4.2). However, on the "Technology & Science" label, the classifier achieves higher recall and higher precision with the human translation. Interestingly, the classifier struggles on both translations with regards to the "Industry & Transport" label with F_1-scores of 0.080 (human translation) and 0.105 (DeepL). Only one of 9 advertisements was correctly labeled in this class and in the human translation 15 advertisements belonging to six different other classes were erroneously labeled as "Industry & Transport". The performance difference is greatest with regards to the "IT" label with 0.427 F_1-score difference. IT job advertisements should, in comparison to other classes, be rather easy to classify. This is due to IT job openings having very specific requirements when it comes to programming languages

French>German			
Translation Version	Raw Human	Original DeepL	
Label	F_1-score	F_1-score	Support
Industry & Transport	0.615	0.545	5
Technology & Science	0.667	0.500	2
Trade & Sales	0.400	0.500	2
Office & Administration	0.286	0.500	5
Financial & Fiduciary Services	0.400	*0.000*	3
Hospitality & Personal Services	0.333	0.500	4
Health	1.000	1.000	1
Teaching & Public Services	0.750	0.769	8
Accuracy	0.567	0.533	30
Macro Average	0.556	0.479	30
Weighted Average	0.539	0.546	30

Table 4.1: Classification results on raw human translations and DeepL-translations from French to German. Italic: scores below 0.25

English>German			
Translation Version	Raw Human	Original DeepL	
Label	F_1-score	F_1-score	Support
Industry & Transport	*0.080*	*0.105*	9
Technology & Science	0.548	0.464	26
IT	0.362	**0.789**	104
Trade & Sales	0.494	0.538	26
Office & Administration	0.435	0.500	31
Financial & Fiduciary Services	0.585	0.720	65
Management & Organisation	*0.103*	0.379	78
Hospitality & Personal Services	0.500	0.800	2
Health	*0.200*	0.462	7
Teaching & Public Services	0.356	0.481	20
Accuracy	0.405	0.590	368
Macro Average	0.366	0.524	368
Weighted Average	0.366	0.585	368

Table 4.2: Classification results on raw human translations and DeepL-translations from English to German. Italic: scores below 0.25, bold: greatest difference

and environments. These specific terms often are capitalised, camel-cased, or in all capitals. Thus, it is my interpretation of this result that the lowercasing of the human translation massively hinders the SJMM classifier's performance in such cases. A very similar issue may occur with the class "Health", which in itself too is a broad field with very specific requirements and terms. The class "Management & Organisation" may suffer from a similar issue, here most likely due to all capital abbreviations. This class was mostly misclassified as "Financial & Fiduciary Services" and as "Office & Administration". As jobs in "Management & Organisation" are responsible for financial aspects of their business and it is a closely related field to "Office & Administration", this is a comprehensible error. In the lower-cased human translation, "Management & Organisation" differs in so far, as it also misclassifies these advertisements as "Trade & Sales". This, again, is a comprehensible error, as "Management & Organisation" may well be responsible for leading employees belonging to "Trade & Sales". This is strengthening my

hypothesis that the lowercasing of e.g. abbreviations is the cause of this even worse performance.

4.2 Analogous Pre-Processing on DeepL-translation

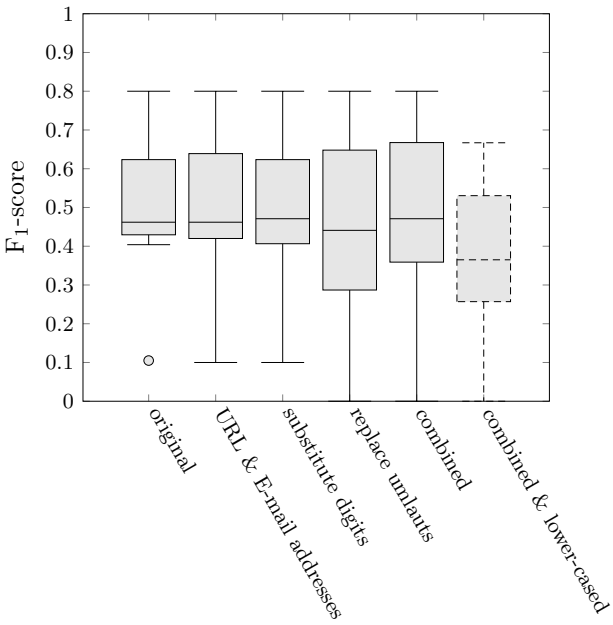

Figure 4.1: F_1-scores on DeepL-translations pre-processed
English > German

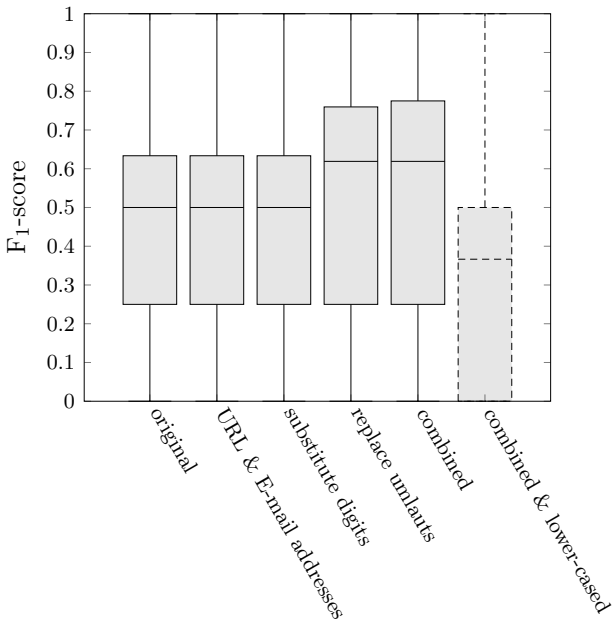

Figure 4.2: F_1-scores on DeepL-translations pre-processed
French > German

As I decided not to use mixed language advertisements due to the scope of this work, the number of advertisements for English sinks to 341 and the French ones to 29.
It must be noted that the differences on the English advertisements are more expressive. This is due to the very low number of advertisements in French. In a first step, I applied the same masking steps as were applied to the human translations. These steps are the masking of e-mail addresses and URLs, of physical addresses, of a person's names, of phone numbers, and setting numerical characters to «0». (See Appendix: Tables: English>German DeepL-translation Processing Experiments for the set of tables showing single steps.)
Masking e-mail addresses and URLs, as well as substituting numerical characters for «0», did only have minor negative effects: The affected classes were "IT", "Trade & Sales", "Office & Administration", "Financial & Fiduciary Services", and "Management & Organisation". With regards to substituting numerical characters, the averaged absolute deltas were < 0.02 and the maximumlabel-specific F_1-score delta was -0.039 for "Management & Organisation". The effect of masking e-mail addresses and URLs was even lower with averaged absolute deltas being < 0.01 and a maximumlabel-specific F_1-score delta of -0.04 for "Trade & Sales". Replacing diacritic Umlauts with their digraphic representation had the largest effect. Whilst the averaged F_1-scores' absolute deltas stayed below 0.05, "Trade & Sales" dropped -0.062 with a support of 25 advertisements and "Health" dropped by -0.262; this is, however, a rare label on this English

set, with only 7 advertisements. Interestingly, combining all three steps did not accumulate these deltas, but rather, for the averaged F_1-score values, it ended up in-between substituting the Umlauts and substituting numerical characters. Forlabel-specific F_1-score values, there were more substantial deltas around -0.1. In the hardest class "Industry & Transport", all steps together caused the classifier to even miss the last advertisements thereof which it previously correctly recognised, resulting in 0 recall.
Finally, I lower-cased the DeepL-translation with all steps applied. As expected from section 4.1, this did have a major negative impact. The weighted average F_1-score dropped by -0.169. Interestingly, the class "Technology & Science", with a support of 25 advertisements, profited from these pre-processing steps, even outperforming the unprocessed translation by +0.13. As mentioned in section 4.1, this is also the one class for which the classifier performed better on the human translation.
The French set only contains 29 advertisements, thus differences in pre-processing produce less but more abrupt changes. Whilst masking URLs and e-mail addresses, and substituting digits did not change the result, replacing umlauts seemingly increases the performance. Due to the low amount of data, this may be coincidental. Results are consistent with the English set when it comes to lower-casing being detrimental though.

English>German: Fully Processed and lower-cased DeepL-translation				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0	0	0	9
Technology & Science	0.475	0.76	0.585	25
IT	0.913	0.228	0.365	92
Trade & Sales	0.395	0.6	0.476	25
Office & Administration	0.387	0.444	0.414	27
Financial & Fiduciary Services	0.477	0.81	0.6	63
Management & Organisation	0.5	0.194	0.28	72
Hospitality & Personal Services	1	0.5	0.667	2
Health	0.25	0.286	0.267	7
Teaching & Public Services	0.167	0.474	0.247	19
racy			0.422	341
average	0.456	0.43	0.39	341
average	0.555	0.422	0.402	341

Table 4.3: Classifier results on fully processed and lower-cased DeepL-translation, English>German

4.3 Unprocessed Human Translation

Due to the results of the previous section, I then also tested the performance on the final cleaned and re-processed human translation. Re-processed here means, I true-cased the translation, substituted the Umlauts, and replaced the Unicode replacement character () with the proper character. At this point, I also had removed the duplicates occurring in these two sets. In the French set, there were two duplicates, in the English set there were nine duplicates. This leaves 27 unique, monolingual French advertisements and in the case of English, there remain 332 thereof. These removals can also change the results of the classification based on the DeepL-translation, thus I re-ran the original DeepL baseline without the duplicates. Whilst the weighted average F_1-score rose by +0.023, the macro average F_1-score fell by the same amount, i.e. the performance difference between classes increased.

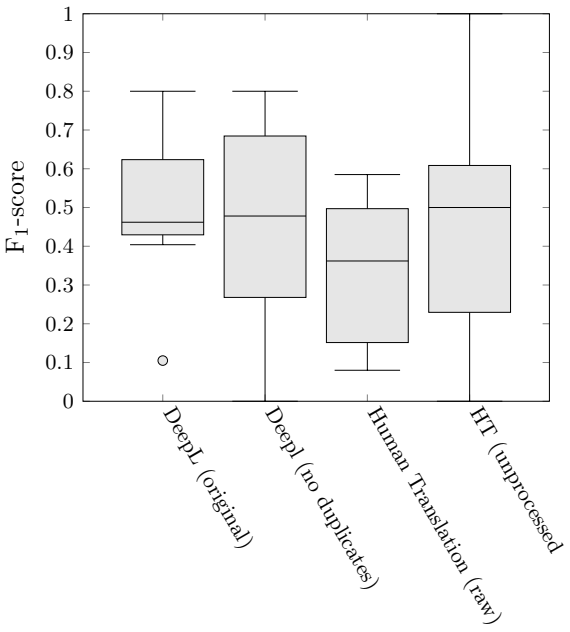

Figure 4.3: F_1-scores on DeepL-translations and Human Translations, with and without duplicates.
English > German

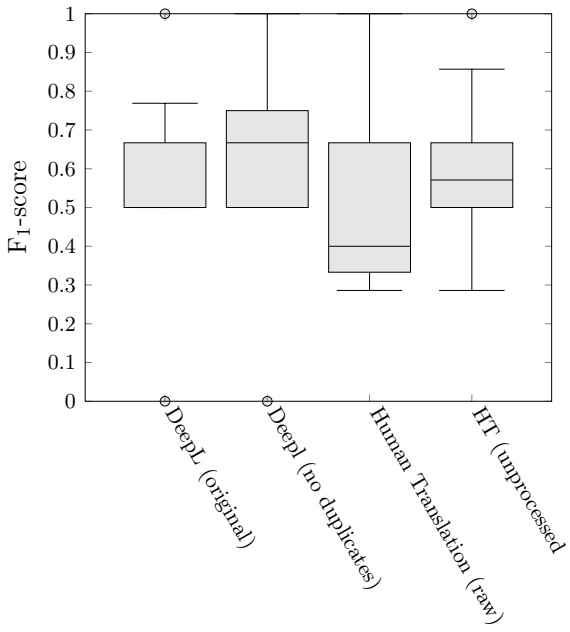

Figure 4.4: F_1-scores on DeepL-translations and Human Translations, with and without duplicates.
French > German

4.3.1 French

For the French set, the re-processing and cleaning have led to a performance increase for the classifier. The weighted average F_1-score increased from 0.539 to 0.62 (+0.081). Of the two duplicate ads, one was classified correctly twice (Industry & Transport) lowering the F_1-score of the class and one was misclassified once (Trades & Sales) but not affecting the F_1-score of the class. All classes not affected by a removal had the same or higher F_1-score after the re-processing.
Compared to the DeepL-translations without duplicates, this result is slightly worse. Accuracy is 3.7% worse (DeepL w/o duplicates: 63%) and the weighted average F_1-score is lower by 0.018 (DeepL w/o duplicates: 0.638).

4.3.2 English

On the English set, the average performance on the human translations rose even steeper, with an increase of the weighted average F_1-score from 0.366 to 0.552 (+0.186) on the re-processed translation. This steep increase further supports the suggested detrimental effect lowercasing had on the results described in section 4.1.
Here, the average performance on the DeepL-translations remains significantly higher. The accuracy on the human translations is 4.8% below that of the DeepL-translations (DeepL w/o duplicates: 60.5%). The weighted average F_1-score is lower by 0.049 (DeepL w/o duplicates: 0.601).

4.4 Best Baseline

The results in the previous section show that the DeepL-translations without duplicates still provides slightly better performance than the human translations. This could be due to the language flattening happening with machine translation (Vanmassenhove et al., 2021), which may ease the automatic classification. Considering that for most advertisements with gold standard labels, only DeepL-translations are available this is good to know. It means that if my translation system outperforms the DeepL baseline, it will be the best translation option available for the current SJMM classifier. This choice was additionally supported by the fact that Gnehm and Clematide (2020) used these translations to explore the efficacy of machine translation as an information transfer method. To establish a baseline on the development sets, I gathered the corresponding DeepL-translations. As mentioned in section 3.1.2, in the Italian development set, either the the DeepL API has failed to translate two of the advertisements or the data was missing from the start. Additionally, in the English development set, 1 advertisement was affected in the same way. In the French development set, all 400 advertisements were translated successfully.

4.4.1 Class Composition of the Gold Standard

As Gnehm and Clematide (2020) have mentioned in their paper, the class distribution of English job advertisements is quite different to those of the French and Italian ones (see tables 4.4 - 4.6 in the next section). I have calculated the percentage of the different classes across the three languages and for the development set, compared to the full set. The random selection preserved the ratios of the classes quite well for the two larger development sets, English (350 elements) and French (400 elements). The Italian development set (74 elements) has larger deviations from the original class ratios. This had to be expected due to the small size.
The most obvious difference is the "Construction" class. Over the full gold standard, only three (0.1%) English job advertisements belong to this class; the development set contains none. The French and Italian sets contain 3.7% and 3.2% "Construction" job advertisements respectively. Similar, but less extreme, ratios can be seen for "Industry & Transport", "Trade & Sales", "Hospitality & Personal Services", and "Health". On the other hand, the more international fields "Technology & Science", "IT", and "Management & Organisation" make up a much larger part of the English set than of the French or Italian sets.

4.4.2 DeepL Baseline for Final NMT Models

As mentioned above in section 4.4, I use the DeepL-translation of advertisements contained in the development set for the new neural machine translation models as a baseline. For comparability, these sets were reduced to the advertisements which (1) remained after cleaning and (2) were successfully translated by all model variants. This leaves 342 English advertisements, 367 French advertisements, and 60 Italian advertisements. Based on the previous work by Gnehm and Clematide (2020), a lower performance due to the different class distribution can be expected on the English set. On the development set, the performance on DeepL-translations of the English set is on average ca. 10% lower than that on that of the French and Italian set.
There is only one label where the precision was higher on the English set than on both of the others: "Management & Organisation". This class is one of those that is massively more frequent

in the English development set (22.6%) than in the French (5.0%) and the Italian (6.9%) ones. Due to very low recall, this does not create as much difference in the F_1-score. On the English development set it is only 0.133 higher than the one on the French development set which itself has low recall for this class, and only 0.019 higher than on the Italian development set. Another class to take note of is "Industry & Transport" with a recall of only 0.154 and a F_1-score of 0.222 in English. This could be caused by the much different ratio of this class in English (4%) compared to French (14.8%) and Italian (18.1%). The last label where there seems to be a connection between ratio and classifier performance is "Teaching & Public Services". In this class, the higher the ratio is, the higher the performance. Due to the low amount of data and especially in Italian, this may also be coincidental though. For the rest of the labels, there is seemingly no connection between the percentage thereof and its classifier performance. I will check my NMT models if the performance on these previously mentioned classes is similar on the English set compared to the French and Italian ones or if this is a DeepL specific issue.

English>German: DeepL-translation Development Set					Class Distribution	
Label	Precision	Recall	F_1-score	Support	Dev Set	Full Set
Industry & Transport	0.4	0.154	0.222	13	4	3
Construction	-	-	-	0	0	0.1
Technology & Science	0.468	0.706	0.562	51	14.9	15.7
IT	0.847	0.769	0.806	65	18.9	17.9
Trade & Sales	0.545	0.75	0.632	16	4.9	5.3
Office & Administration	0.472	0.63	0.54	27	8	8.1
Financial & Fiduciary Services	0.631	0.745	0.683	55	15.8	16.4
Management & Organisation	0.815	0.282	0.419	78	22.6	22
Hospitality & Personal Services	0.667	0.667	0.667	3	0.9	0.9
Health	1	0.429	0.6	7	2	2.4
Teaching & Public Services	0.444	0.741	0.556	27	8	8.2
Accuracy			0.599	342		
Macro Average	0.629	0.587	0.569	342		
Weighted Average	0.657	0.599	0.585	342		

Table 4.4: Best classifier results of the DeepL-translation to be used as a benchmark and class distribution on the gold standard in percent, English > German

French>German: DeepL-translation Development Set					Class Distribution	
Label	Precision	Recall	F_1-score	Support	Dev Set	Full Set
Industry & Transport	0.755	0.769	0.762	52	14.8	14
Construction	0.857	0.545	0.667	11	3.3	3.7
Technology & Science	0.529	0.75	0.621	24	6.5	7.9
IT	0.875	0.824	0.848	17	4.5	4.5
Trade & Sales	0.767	0.647	0.702	51	13.5	12.3
Office & Administration	0.71	0.468	0.564	47	12.5	10.8
Financial & Fiduciary Services	0.842	0.744	0.79	43	11.3	10.6
Management & Organisation	0.261	0.316	0.286	19	5	7.5
Hospitality & Personal Services	0.926	0.676	0.781	37	11	8.8
Health	0.742	0.852	0.793	27	7.5	7.8
Teaching & Public Services	0.547	0.897	0.68	39	10.3	12.2
Accuracy			0.692	367		
Macro Average	0.71	0.681	0.681	367		
Weighted Average	0.724	0.692	0.694	367		

Table 4.5: Best classifier results of the DeepL-translation to be used as a benchmark and class distribution on the gold standard in percent, French > German

Italian>German: DeepL-translation Development Set					Class Distribution	
Label	Precision	Recall	F_1-score	Support	Dev Set	Full Set
Industry & Transport	0.5	0.4	0.444	10	18.1	13.3
Construction	0.667	0.667	0.667	3	4.2	3.2
Technology & Science	0.429	0.6	0.5	5	9.7	7.5
IT	1	1	1	1	2.8	2.5
Trade & Sales	0.714	0.625	0.667	8	11.1	13.8
Office & Administration	0.667	0.571	0.615	7	9.7	12.3
Financial & Fiduciary Services	1	0.8	0.889	5	9.7	10.3
Management & Organisation	0.333	0.5	0.4	4	6.9	7.9
Hospitality & Personal Services	0.857	0.857	0.857	7	11.1	12.0
Health	1	0.714	0.833	7	12.5	9.2
Teaching & Public Services	0.333	0.667	0.444	3	4.2	7.9
Accuracy			0.633	60		
Macro Average	0.682	0.673	0.665	60		
Weighted Average	0.681	0.633	0.647	60		

Table 4.6: Best classifier results of the DeepL-translation to be used as a benchmark and class distribution on the gold standard in percent, Italian > German

5 Model Variants

5.1 Generic Base Model

The Generic Base Model is a transformer model (Vaswani et al., 2017) trained on the corpora described in section 3.1 on Sockeye 2.0 (Hieber et al., 2020). Due to the scope of this work and the available resources, I do not have the possibility of experimenting with hyperparameters. Thus, I mainly used Sockeye's default values. The config file used can be found in the Appendix of this work. For this first step, I did not yet concatenate any sentences, as there is next to no possibility to differentiate consecutive segments from independent segments in the corpora used. This model was trained on 12.2 million segments for 500'000 updates with a vocabulary size of 32'794, taking 274 hours on two NVIDIA GeForce GTX TITAN X GPUs.

5.2 Alpha Model Variants

For the alpha models, there are four main variants, which are described in the following subsections. I checked for over-fitting and/or catastrophic forgetting by using the generic development set as the validation set here too. The performance on the in-domain data was checked afterwards.

5.2.1 Alpha Variant 1: Fine-Tuning on Single Segments

This first variant is trained on single segments only. The domain adaption used is pure fine-tuning on the human translations and the job titles. Whilst pure fine-tuning has strong tendencies towards over-fitting and catastrophic forgetting, it requires comparatively very little computation time. Due to the low amount of in-domain data, it proved difficult to get enough data input in order to adjust the model's weights properly. The first attempt ended with a complete loss of the model due to the adapted learning rate being too high and the amount of input being too low. Interestingly, the default learning rate of Sockeye proved effective both for the generic training and the domain adaption training. I trained the model for a single epoch to avoid over-fitting. This took 15 minutes on two NVIDIA GeForce GTX TITAN X GPUs. Over this one epoch, training perplexity dropped from 51.84 to 25.14. Meanwhile, the validation perplexity rose from the base model's 14.9 to 17.34, not indicating over-fitting yet. Running a second epoch lead to over-fitting already.
This first variant mainly served as a comparison to the second variant.

5.2.2 Alpha Variant 2: Fine-Tuning on Two Concatenated Segments

As mentioned in the previous section, this model is the same in terms of training method but differs in the way the in-domain data is presented. Here, every two consecutive segments that belong to the same job advertisement are concatenated. The border between segments is indicated by the special token <seg>. This was both a test both

(1) for the performance compared to single segments, as well as

(2) for the feasibility of only introducing these concatenated segments at domain adaption time.

The results regarding these two points were critical for the further variants, as the scope of this work was very tight. Point (1) will be discussed in section 6 Alpha Model Results. The result of point (2) was highly promising. Even with the tiny amount of in-domain data, the model already learned to split the content into two parts when a <seg> token appeared. As a result, it reliably created two segments out of said input. Even with the amount of cleaning performed on the in-domain data, there still were very few segments containing only erroneously segmented symbols. In such cases, the model created only one segment. In this circumstance, this was indeed a quite desirable, easy to handle solution to such noise. Thus, this approach is feasible even with very little data.
In the following example, the question mark symbol was used to mark items in a bulleted list (BPE removed for better readability):

(5.1) source: "<2de> ? <seg> To develop and implement the European Member and Partnership acquisition strategy and the European regional marketing plan"

(5.2) translation: "sie entwickeln und implementieren die europäische Annahmestrategie und Partnerschaftsplanung und den europäischen regionalen Marketingplan ."

Again, this model was domain-adapted for one epoch only. This took 17 minutes on two NVIDIA GeForce GTX TITAN X GPUs. It has to be kept in mind though, that, as segments are concatenated, they are in effect doubled. Training perplexity was very similar to the single segment model starting at 51.2 and lowering to 24.0 in the end. Validation perplexity rose from the base model's 14.9 to 17.9. Merely based on these numbers, a slight improvement over the previous model could be expected. For the results, see section 6. It was interesting to see, that the validation perplexity on the generic development set did not rise much higher than for the previous model, even though the domain-specific data was doubled.

5.2.3 Alpha Variant 3: Fine-Tuning Supported by Generic Data

As discussed in the previous two sections, catastrophic forgetting is a highly destructive issue with pure fine-tuning. These next two variants aim to reduce this. As concatenating two segments proved useful, I continued using this method for these variants.
In order to combat catastrophic forgetting, this variant is trained on a combination of over-sampled in-domain data and generic data from the base model. The generic data added will re-activate the previously learnt weights, thus keeping them relevant which in turn avoids catastrophic forgetting to an extent. I over-sampled the concatenated human translations by a factor of three. Meanwhile, the job title terms were only used once. I then added one generic segment per in-domain segment to the training data. Again, I ran this for one epoch only. Owing to the much larger data set, this single epoch took 44 minutes on two NVIDIA GeForce GTX TITAN

X GPUs.
As expected when adding training data previously used in the pre-trained model, the starting point for training perplexity is lower at 32.53. After one epoch it had dropped to 17.12; slightly lower than the previous two variants. Validation perplexity was at 14.48, even improving over the pre-trained model. Keep in mind that due to the scope of this work, the pre-trained model was not trained to convergence. This is a first interesting result as it shows an opportunity for optimisation in circumstances where the available computing time is limited.

5.2.4 Alpha Variant 4: Fine-Tuning Using Elastic Weight Consolidation

In this variant, catastrophic forgetting is combated by using Elastic Weight Consolidation (Thompson et al., 2019). As mentioned in section 2, this means that the Empiric Fisher Information is calculated on the weights of the pre-trained model. This allows the machine learning algorithm to focus on adapting weights which are less impactful to the performance in the generic domain. Whilst lower computational needs based on the lower amount of input data can be expected, there is also slightly increased computational need, as the Empirical Fisher Information of the model's parameters has to be checked. This model contains 60'927'002 trainable parameters. However, at the time of writing this work, Sockeye only provides a draft implementation of Elastic Weight Consolidation where GPU support is not yet properly enabled. The most time-efficient solution was to create a workaround by hard coding one GPU to be used. I used only the in-domain data and trained the model for three epochs. This is comparable to the over-sampling by a factor three used in the previous variant. Training took 49 minutes on a single NVIDIA GeForce GTX TITAN X GPU; only five minutes more than the previous version (Alpha Variant 3: Fine-Tuning Supported by Generic Data) with only half the GPU resource used.
Training perplexity lowered from 59.25 to 13.53. Validation perplexity rose significantly, however, ending at 69.12. This suggests that this model already reached the decay phase of the generic domain. Based on training perplexity, a similar state to the previous versions' end state was reached around 2.5 epochs into training. Due to the very low amount of data available, it was not feasible to create more validation checkpoints in order to find the optimal state. Additionally, the scope of this work did not allow for much experimentation trying to adapt hyper-parameters or the training length.

6 Alpha Model Results

6.1 Generic Base Model

6.1.1 English

For English, the Generic Base Model achieved a BLEU-score of 28.98 (against DeepL). Comparing the classification performance, as expected, this generic model with limited training data and training time is worse than DeepL. The two translations differ by 7.4% in accuracy and the weighted average of the F_1-score differs by 0.068. I expect this difference to be surmountable using domain adaptation.
As I have added the European Central Bank corpus to the generic Data, I expected a smaller gap on the class "Financial & Fiduciary Services". This class is indeed only trailing by 0.037 in F_1-score on this model. The class "Technology & Science" is also worth taking note of. It is one of the more frequent classes in this set, and in this class, the classifier already performs better on the Generic Base Model, owing to higher precision (+0.32).

6.1.2 French

On the French development set, the Generic Base Model achieved a BLEU-score of 27.39 (against DeepL). Regarding the classification performance, the difference is larger compared to the English set. This model's accuracy is 11% below DeepL and the weighted average of the F_1-score is lower by 0.123.
Notably, this model failed to classify any advertisement belonging to "Construction" correctly. "Financial & Fiduciary Services" is the second-best performing class on this set, behind "Health", with slightly higher recall on this model (+0.47).

6.1.3 Italian

The Generic Base Model has a BLEU-score of 28.95 (against DeepL).
The Italian set presents the highest classifier performance difference between this model and DeepL. The accuracy is 16.5% lower on this model. On single classes, the amount of data is very low. One point to note, though, is the 0 score on the classes "Construction" (like on the French set) and "Management & Organisation".

6.2 Alpha Variant 1: Fine-Tuning on Single Segments

As previously mentioned, this model mainly serves as a comparison between single segment training and translation and concatenated segments training and translation. (I.e. comparing

it to alpha variant 2.) It is expected to be outperformed by Variants 3 and 4. This expectation does need to be confirmed though.

6.2.1 English

Compared to the Generic Base Model, the BLEU-score increased to 33.74 (+4.79). The classifier performance increases on all but two classes. These two classes are the rarest ones of the English development set: "Hospitality & Personal Services" (3 advertisements) where performance remains equal, and "Health" (7 advertisements) on which performance slightly drops. This drop is caused by lower recall (-0.143), whilst precision rose (+0.125).
With regards to the DeepL Baseline, the difference in accuracy shrank to -3.6% (+3.8%) and the weighted average F_1-score differs by -0.033 (+0.027).

6.2.2 French

On the French set, the BLEU-score sinks to 26.69 (-0.64) compared to the Generic Base Model. The averaged classification values, accuracy and weighted average F_1-score, remain within 0.01 of the Generic Base Model. Compared to the DeepL Baseline, this variant's accuracy is lower by -11.2% at 58% and its weighted average F_1-score is lower by -0.113 at 0.581. Of the eleven "Construction" advertisements, two are now correctly classified (previously 0).

6.2.3 Italian

Finally, on the Italian set, the BLEU-score sinks to 27.78 (-1.17) compared to the Generic Base Model. However, accuracy rises by 8% to 54.8% and the weighted average F_1-score rises by 0.075 to 0.554. The two 0 score classes do not improve. "Technology & Science" loses 0.2 recall as one more of its five advertisements is misclassified.
These results are highly remarkable as there are no Italian human translations. The only in-domain data for Italian are the job title terms.

6.3 Alpha Variant 2: Fine-Tuning on Two Concatenated Segments

The very similar perplexity values measured during training and validation of this model are mirrored by small performance differences between this model and alpha variant 1.

6.3.1 English

Compared to alpha variant 1, the BLEU-score is slightly lower at 33.2 (-0.54) on the English development set. The classification performance remains almost the same. It would seem, that on the English set the concatenation of two consecutive sentences does not provide improvement.

6.3.2 French

On the French development set, the BLEU-score of this variant at 26.71 is just slightly higher than for alpha variant 1 (+0.02). Opposed to the English set, here classification performance does increase on all but three classes, resulting in an improved average accuracy of 0.611 (+3.1%) and the weighted average F_1-score is higher by 0.029 at 0.6. The performance on the "Construction" label remained equal. The classes with lower performance were "Office & Administration", as well as "Management & Organisation".

6.3.3 Italian

The BLEU-Score on the Italian development set, compared with version 1, rose by 1.7 to 28.41. The average accuracy improved slightly (+1.9%), but more importantly, the Construction label is now found once.

From these results it can be said, that context extension by one sentence, on average, does not deteriorate the results, but can provide a slight improvement on this task. Thus I continue this work using concatenated data. In a continuation of this work, more extensive tests with larger context extension could prove beneficial.

6.4 Alpha Variant 3: Fine-Tuning Supported by Generic Data

As this variant reached a lower perplexity than the two previous variants during training and even improved the validation perplexity over the Generic Base Model, I expect higher performance from this model.

6.4.1 English

This variant's translation achieved a BLEU-score of 34.24, beating the previous best by 0.52. As expected by the perplexity values and the BLEU-score, there is also an improvement in the classifier's performance. This variant beats the previous best by +3% accuracy and +0.032 weighted average F_1-score. This also means that this model is almost on-par with the DeepL baseline. On accuracy it is -0.5% below the performance on the DeepL-translation; on the weighted average F_1-score it lacks merely 0.001. Keeping in mind that this performance is based on a small generic model which is domain-adapted with only 332 English advertisements, or 20166 in-domain segments, and job title terms, this is an impressive development.

6.4.2 French

On the French set, the BLEU-score of this variant is 27.88. This is an improvement over the previous best of +1.17. Improvement on the average classifier scores is lower. Accuracy rises by 1%, the weighted average F_1-score by 0.02. The gap to the performance on the DeepL-translation

is as follows: On accuracy, this variant is 7.1% worse. The weighted average F_1-score is 0.072 points lower on this variant. As there only were 27 unique advertisements, or 350 segments, of French human translation, a lesser development compared to English was to be expected.

6.4.3 Italian

The BLEU-score of this variant on the Italian development set is 30.15. This is an improvement of 1.74 points over the previous best. Whilst the score on the "Construction" label returns to all zeroes again, accuracy and weighted average F_1-score rise significantly. The accuracy of 63.3% is an improvement of 7.4% over the previous best. The weighted average F_1-score improved by 0.071 points to 0.635. This means, that, even though there is no in-domain data besides the job title terms for Italian, the performance on the Italian set is the highest. It has to be kept in mind though, that the Italian development set is small with 60 advertisements. This result may be coincidental. Nonetheless, it is an impressive show of the possibilities offered by multilingual models.

This variant looks highly promising for beating the performance on the DeepL-translations once more in-domain data is available.

6.5 Alpha Variant 4: Fine-Tuning Using Elastic Weight Consolidation

Seeing from the perplexity values that this Model already had issues with catastrophic forgetting due to the low amount of in-domain data, it will be interesting to see how this affects the classifier's performance.

6.5.1 English

The high evaluation perplexity is also reflected in the BLEU-score: With 30.77 points, it is the lowest by 2.43 points. The classifier performance, however, is higher than on the two raw fine-tuning variants (alpha variants 1 and 2) but lower than alpha variant 3. On the label "Trade & Sales", one of the rarer labels in the English set, it even achieved the highest F_1-score out of all variants.

6.5.2 French

For the French set, the results are analogous. With a BLEU-score of 22.54, it is the lowest of any variant, whilst the F_1-score is between alpha variants 2 and 3. Again, "Trade & Sales" performance is the best of any variant here.

6.5.3 Italian

On the Italian set, this variant presents differently. With Elastic Weight Consolidation, the BLEU-score is massively lower at 21.23 (-6.45). This is also reflected in the classification performance. It is just slightly better than alpha variant 1 (accuracy +0.002, weighted average F_1-score +0.003).

6.6 Alpha Variants Conclusion

Based on these results, it is clear that Alpha Variant 3: Fine-Tuning Supported by Generic Data is performing the best on this task. An overview over the four variants can be seen in the following box-plots of their weighted average F_1-scores per label:

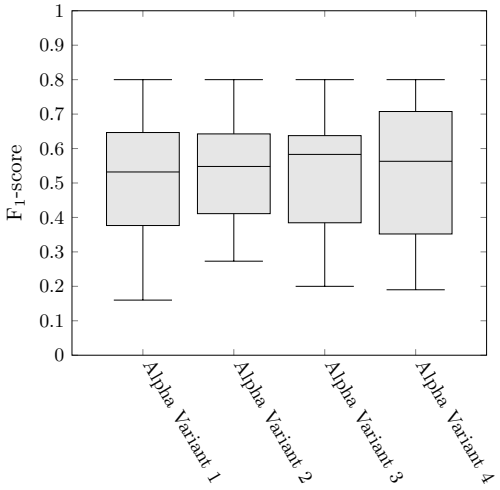

Figure 6.1: Unweighted F_1-scores on Alpha Variants
English > German

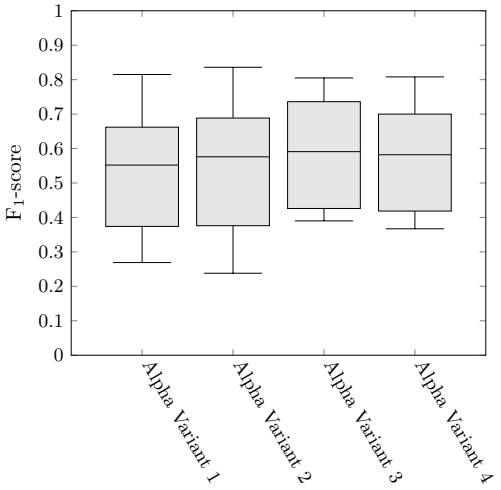

Figure 6.3: Unweighted F_1-scores on Alpha Variants
French > German

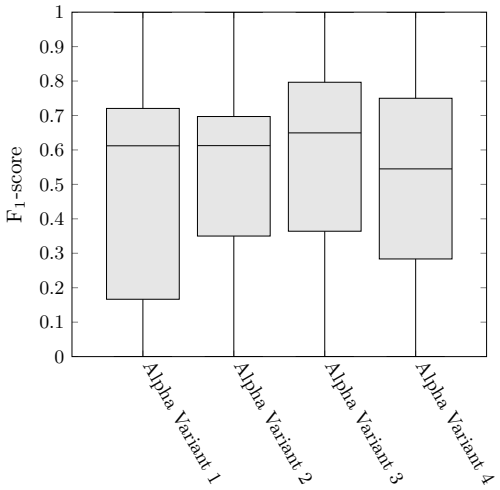

Figure 6.2: Unweighted F_1-scores on Alpha Variants
Italian > German

6.7 Constrained Decoding

Direction	Unconstrained	Constrained	Support
French > German	0.620	0.614	367
Italian > German	0.635	0.651	60

Table 6.1: Weighted average F_1-score compared between constrained and unconstrained decoding. Based on Alpha Variant 3: Fine-Tuning Supported by Generic Data. (See 5.2.3)

As key words, such as job titles, should help the classifier massively, and the in-domain data is heavily limited, I also considered constrained decoding for job titles and education levels. As the scope of this work does not allow me to develop extensive pre- and post-processing modules, I kept this approach simple, with direct matches to the terminology only. This means that morphology in the target segment is ignored which hampers performance.
Using the job title terms and the education level abbreviations, performance sank on all models. I interpret this such that, due to the job title terms in the training data, my model is already able to translate job titles correctly. It is possible that the missing morphology forces the translation into a worse performing probability range for the rest of the tokens to be used around the constraint. One example here-for would be a case where the constraint applies to a direct or indirect object of a phrase but due to lacking morphology, the system builds the translation with the constraint as the phrase's subject.
In the DeepL-translations, mainly education level abbreviations were problematic. E.g. the French «CFC» (Certificat fédéral de capacité), meaning federal certificate of proficiency (my translation), was interpreted as Chlorofluorocarbure (Chlorofluorocarbon). This led me to reduce the terminology to education level abbreviations. These are also not affected by morphology. The only set on which average values improved was the Italian one, based on one more advertisement classified correctly which is not significant. Therefore, I decided not to further pursue constrained decoding in my experiments.

7 Synthetic Data Produced

7.1 Back-translation from German

Based on these results discussed in the previous section, I decided to use alpha variant 3, domain adaptation with generic support (see 5.2.3), without constraint based decoding for the back-translations. For the back-translation, I used exclusively advertisements that do not yet have a gold standard label, in order to keep as much test data as possible. To start out with, I selected the first and last 50'000 advertisements from a folder which does not contain any advertisement with a label. These first advertisements are from 2014, the last ones from 2018. Each of the source files contains ca. 100'000 advertisements. Thus, by selecting 50'000 advertisements, I do not risk double sampling job advertisements where the position still is listed at the next data collection date. There still is a small risk of double sampling though, since these advertisements stem from different job listing sites. This resulted in ca. 1.3 million segments per language pair.

7.2 Forward-Translation to German

Using alpha variant 3, I forward-translated 200'000 segments from French and English. From Italian, I translated 166'988 segments; these are all marked Italian advertisements which have no gold-standard label.

8 Models Using Synthetic Data

If not explicitly mentioned otherwise, for the following experiments, I based the domain-adaptation off of the baseline model (see Section 5.1) and limited training to be unidirectional into German as this is these models' only intended use.

8.1 Variant 1: Using All Back-translations

8.1.1 Variant 1.1: Supported by Half of the Generic Data.

For this training, I used the same in-domain data as in the alpha models, enriched with all 100'000 back-translated advertisements. To support this, I added half of the generic data used for the generic base model. By using half of the generic data, there is roughly the same amount of out-of-domain and in-domain Data in the training material. This results in improved performance over the best alpha model variant (see Section 5.2.3) in English and French.

8.1.1.1 English

The English set reaches a BLEU-score of 36.31 (against DeepL), thus surpassing the previous best by +2.07 points. Compared to the previous best, this model reaches an accuracy of 60.8% (+1.4%) and a weighted average F_1-score of 0.602 (+0.018). On a per label basis, this model is:

- under-performing on two labels ("Office & Administration", and "Teaching & Public Services") by -0.041 and -0.019 respectively.
- on par for three labels ("Trade & Sales", "Management & Organisation", and "Hospitality & Personal Services").
- slightly out-performing three labels ("Technology & Sciences", "IT", and "Financial & Fiduciary Services") by +0.03, +0.038 and +0.006 respectively.
- strongly out-performing two labels ("Industry & Transport", and "Health") by +0.185 and +0.167 respectively.

This classifier performance, for the first time, is also higher than the DeepL-baseline with +0.7% accuracy and +0.017 weighted average F_1-score. This is remarkable and encouraging, as the model which has produced these back-translations was domain-adapted on only a few hundred human-translated advertisements and a set of job titles and education level abbreviations.

8.1.1.2 French

Improvement on the French set is lower. With a BLEU-score of 28.38 (against DeepL), it outperforms the previous best by +0.5 points. Accuracy is at 62.9% (+0.8% over the previous best) and the weighted average F_1-score reaches 0.631 (+0.011 over the previous best). A look

at the weighted average F_1-score of the different labels shows that only six labels increased in performance, whilst five dropped in performance. Compared to the DeepL-baseline it is still under-performing by -6.9% accuracy and by -0.063 weighted average F_1-score.
It would seem, that fewer human-translated advertisements in domain adaption also lead to lower performance gains on back-translations.

8.1.1.3 Italian

On the Italian set, where no human-translated advertisements were available, the BLEU-score with 29.55 (against DeepL) is lower than the previous best by -0.6 points. As can be expected based on this, accuracy at only 58.3% (-5%) and the weighted average F_1-score at 0.588 (-0.47) are lower also.
It would seem, that the previous success of transfer learning cannot be extended to the production of performance-boosting back-translations.

8.1.1.4 Extended Domain-Adaptation

As with the alpha models, extending the domain-adaptation past 1 epoch lowers the performance on all three languages. It is my belief that the errors remaining in the un-processed human translations and the translationese and errors of the back-translations overshadow the usefulness of said data at this point.

8.1.2 Variant 1.2: Supported by a Quarter of the Generic Data.

In a next step, I halved the out-domain data supporting the domain-adaptation in order to establish its impact. The rest of the data remains equal. After the evaluation of this model, there is an unfortunate situation. The best classifier performance for each of the three source languages stems from a different model:

- Best English: Model using all back-translations, supported by half the out-domain data. (Better than DeepL-baseline)
- Best French: Model using all back-translations, supported by a quarter of the out-domain data.
- Best Italian: Alpha model variant 3.

8.1.2.1 English

The BLEU-score is 35.45, -0.86 points compared to the previous model. And whilst the BLEU-score is higher than the best alpha model, the classifier performance on this model's translation is lower than the best alpha model. Accuracy at 57.9% is -2.9% lower compared to the previous model using half the generic data. The weighted average F_1-score is even lower at 0.559; a reduction of -0.043.
It should be noted, that when comparing alpha model variant 3 with this model, the BLEU-Score fails as a performance predictor for the first time in this work.

8.1.2.2 French

For the French set, the BLEU-score is at 28.21 (-0.17 compared to the previous model). Despite having a slightly lower BLEU-score than the previous model using half the out-domain data, the classifier performance increases to a new best. Accuracy reaches 65.4% (+2.5%) and the weighted average F_1-score is at 0.653 (+0.022).

8.1.3 Italian

The BLEU-score on the Italian set is 30.21 (+0.06 over the previous best; +0.66 over the previous model). Despite this, the classifier performance only slightly increases when considering the weighted average F_1-score over the previous model; the accuracy remains the same.

8.2 Adding Forward-translations

Based on the findings of Bogoychev and Sennrich (2019), I added the forward-translations to the training data used for the previous section. This may improve results over only using back-translations.

8.2.1 Variant 2.1: Forward-translations Added to Variant 1.1

Variant 1.1's training data contains half of the out-domain data. After adding the forward-translations a clear rise in BLEU-scores can be observed, with slight classifier performance improvements for French and Italian at a slight cost of English classifier performance.

8.2.1.1 English

This model set a new highest BLEU-score for the English set of 37.16 (against DeepL) (+0.85 points over variant 1.1). Accuracy is at the level of alpha variant 3 (59.4%) and the weighted average F_1-score at 0.589 is slightly better than that, but still below variant 1.1.

8.2.1.2 French

In French, this model also achieves a new highest BLEU-score of 29.96 (+1.58 points over variant 1.1). Additionally, accuracy rises compared to variant 1.1, but remains below variant 1.2, with 63.5%. The weighted average F_1-score equals variant 1.1 at 0.631.

8.2.1.3 Italian

And finally, on the Italian set, there is another new highest BLEU-score of 32.47 (+2.26 over variant 1.2). The accuracy on this translation rises to 60%, thus being higher than both variants 1.1 and 1.2 but below alpha variant 3. The same is true about the weighted average F_1-score at 0.606.

8.2.2 Variant 2.2: Forward-translations Added to Variant 1.2

Whilst, similarly to variant 2.1, the BLEU-score rises on all languages by +1.6 or more points, classifier performance on French and English translations sinks compared to variant 1.2. On the Italian translations, on the other hand, a new highest performance was achieved, which surpasses the DeepL-baseline.

8.2.2.1 English

The BLEU-score of 37.05 is +1.6 points better than version 1.2. The classifier performance sinks, however, to 57.3% accuracy and a weighted average F_1-score of 0.563. Thus, adding forward-translations lowered the classifier accuracy by -1.2% and the weighted average F_1-score by -0.011.

8.2.2.2 French

The BLEU-score of 30.04 (+1.83 over version 1.2) is a new highest value. Classifier performance sinks though, with an accuracy of 64% (-1.4%) and a weighted average F_1-score of 0.636 (-0.017).

8.2.2.3 Italian

Compared to variant 1.2, the BLEU-score rises by +2.26 points to 32.47, on par with variant 2.1. Whilst previously the Italian models including synthetic parallel data had lower classifier performance than the alpha variant 3, this model's translations achieve 65% accuracy (+1.7% over the previous best; +1.7 over the DeepL baseline) and a weighted average F_1-score of 0.647 (+0.012 over previous best; on par with DeepL-baseline).
On a per label basis, this model:

- under-performs compared to DeepL on six labels ("Construction", "Technology & Science", "IT", "Financial & Fiduciary Services", "Management & Organisation", and "Health"). Notably, this model scores a zero F_1-score on "Management & Organisation" with four advertisements present.
- is on par with DeepL on the label "Hospitality & Personal Services", the fourth most common label in both the full gold standard and this set.
- outperforms compared to DeepL on four labels ("Industry & Transport", "Trades & Sales", "Office & Administration", and "Teaching & Public Services"). The first two of these labels are the most common classes in the full gold standard.

Once again, it must be recognised that this is a small data set with only 60 advertisements distributed over 11 classes. Considering the performance documented in the previous sections and this low amount of data, this result may be entirely coincidental.

8.3 Variant 3: Elastic Weight Consolidation

Seeing as this method had difficulties with the very low amount of data available for the alpha variants, I decided to test it again using all available synthetic data. For French and Italian,

the performance of EWC was similar to variant 1.1; for English, it was worse. With regards to BLEU-score, on each language, it was more than 2.5 points below the worst model out of variants 1.1, 1.2, 2.1, and 2.2.

8.4 Comparing Variants across Languages

At this point, these are the best models per language:

- English: Variant 1.1
- French: Variant 1.2
- Italian: Variant 2.2 / Alpha Variant 3

In practice, optimally, only one model would be in use. As a different model provided the highest classifier performance for each language, it is also interesting to see, which model performs the best across all languages. Averaging the raw accuracy and weighted average F_1-scores per model over all languages, variant 2.2 would earn the highest score. However, support has to be taken into account. When weighing the calculation with the number of advertisements classified per language, variant 1.1 achieves the highest combined weighted average F_1-score (0.6147) and the second-highest combined accuracy (61.60%); variant 1.2 achieving the highest combined accuracy (61.78%) and second-highest combined weighted average F_1-score (0.6133). Variant 1.1 has the highest classifier performance for English, whilst variant 1.2 has the highest classifier performance for French. Variants 2.1 and 2.2 using forward-translations are interesting. Whilst they have higher BLEU-scores and variant 2.1 outperforming variant 1.1 on two of three languages, losing performance on the third language means, it none the less has lower combined scores.

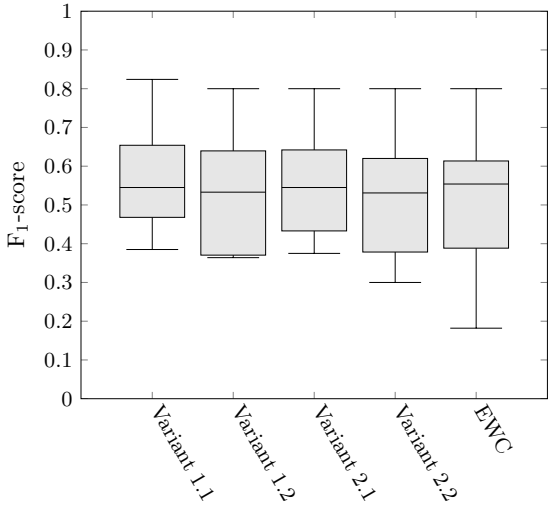

Figure 8.1: Unweighted F_1-scores on Variants using synthetic parallel data on the development set
English > German

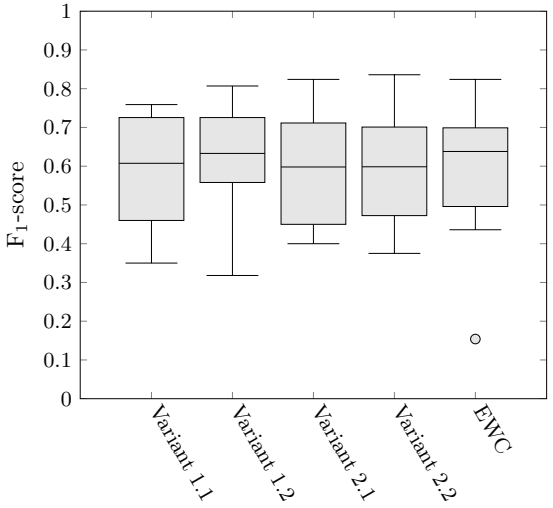

Figure 8.2: Unweighted F_1-scores on Variants using synthetic parallel data on the development set
French > German

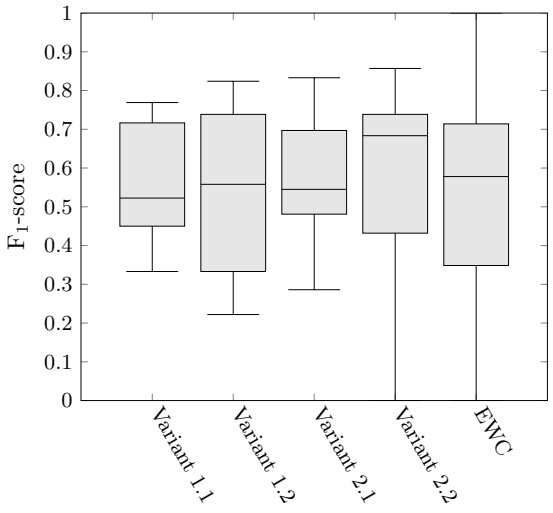

Figure 8.3: Unweighted F_1-scores on Variants using synthetic parallel data on the development set
Italian > German

8.4.1 Result Validation on the Test Set

8.4.1.1 Class Distribution on the Test Set

As the selection for the development set and the test set was random, the class distribution differs in these two sets and also compared to the full set. Thus, during the interpretation of the results, one has to keep in mind the number of advertisements supporting them. The following table shows the distribution differences on the different sets and languages:

Percentage of Advertisements per Class in the Gold Standard									
	English			French			Italian		
Label	Dev Set	Test Set	Full Set	Dev Set	Test Set	Full Set	Dev Set	Test Set	Full Set
Industry & Transport	4.0	1.5	3.0	14.8	15.3	14.0	18.1	16.7	13.3
Construction	0	0.3	0.1	3.3	3.5	3.7	4.2	1.7	3.2
Technology & Science	14.9	16	15.7	6.5	10.2	7.9	9.7	6.7	7.5
IT	18.9	18.3	17.9	4.5	4.3	4.5	2.8	3.3	2.5
Trade & Sales	4.9	6.4	5.3	13.5	13.2	12.3	11.1	18.3	13.8
Office & Administration	8.0	7.3	8.1	12.5	9.4	10.8	9.7	10	12.3
Financial & Fiduciary Services	15.8	16.3	16.4	11.3	8.9	10.6	9.7	6.7	10.3
Management & Organisation	22.6	20.9	22.0	5.0	8.6	7.5	6.9	6.7	7.9
Hospitality & Personal Services	0.9	1.2	0.9	11.0	7.5	8.8	11.1	13.3	12.0
Health	2.0	4.4	2.4	7.5	8.1	7.8	12.5	6.7	9.2
Teaching & Public Services	8.0	7.6	8.2	10.3	11	12.2	4.2	10	7.9

Table 8.1: Distribution of job classes in gold-standard data across all Sets in percent

For English, there are four classes that differ significantly in their distribution between the development and test sets; these being "Industry & Transport", "Construction" (missing from the development set), "Trade & Sales" and "Health".
Looking at the French sets, there are five such classes: "Technology & Science", "Office & Administration", "Financial & Fiduciary Services", "Management & Organisation", and "Hospitality & Personal Services".
In the case of Italian, six classes are distributed significantly differently between the two sets.

These are "Construction", "Technology & Science", "Trade & Sales", "Financial & Fiduciary Services", "Health", and "Teaching & Public Services". As Italian has much fewer advertisements, more change in the distribution was to be expected.

8.4.1.2 Results on the Test Set

For English, the results are highly similar to the development set except for BLEU-scores being lower than on the development set by more than 10 points (Variant 1.1: 25.53, Variant 1.2: 25, Variant2.1: 26.25, Variant 2.2: 26.01). Seeing as most classes have a similar distribution, this was to be expected. One difference impacting mainly the macro average of the F_1-score is the one "Construction" add in the Test set, which was misclassified on all variants, including DeepL. Variant 1.1 which performed best on the development set, is outperformed by variant 2.1 on the test set. This is mainly due to higher classifier performance on the most frequent classes "Technology & Science" (F_1-score +0.041, 16%), "IT" (F_1-score +0.02, 18.3%), "Financial & Fiduciary Services" (F_1-score +0.064, 16.3%), and "Management & Organisation" (F_1-score +0.064, 20.9%). With respect to these two variants, this is directly inverted to the results on the development set. This result shows the difficulty of comparing multiple, closely related models on different sets. An interesting point to take note of is that on the test set variants 1.2 (+1.9% accuracy), 2.1 (+0.9% accuracy), and 2.2 (+0.9% accuracy) all outperform DeepL.
In the case of French where there are more distribution differences, the accuracy on the test set is on average 7.5% higher than on the development set. This is mainly due to large performance gains on two classes which performed worst on the development set: "Technology & Science" and "Management & Organisation". Comparing the two best variants on both sets, Variant 2.1 on the development set and Variant 1.1 on the test set:

- "Technology & Science": F_1-score +0.053
- "Management & Organisation": +0.359 (at 0.656 more than doubling the classifier's performance)

Another interesting class is "Health", already achieving an F_1-score of 0.807 on variant 2.1 on the development set, on the test set it achieves an almost perfect F_1-score of 0.968 with perfect recall on 30 advertisements.
On the two French sets, there is one consistency: the best performing models are trained without forward-translation. The BLEU-scores of the test set are ca. 2 points lower than on the development set (Variant 1.1: 26.28, Variant 1.2: 26.62, Variant2.1: 28.76, Variant 2.2: 28.55). Compared to DeepL, the best model on the test set, Variant 2.1 is 1.6% lower in terms of accuracy and 0.016 points lower in terms of weighted average F_1-score.
Results on the Italian sets are similar to the French ones with even bigger classifier performance differences; BLEU-scores are at a similar level on this set as on the corresponding development set (Variant 1.1: 29.83, Variant 1.2: 29.78, Variant2.1: 30.66, Variant 2.2: 30.97). These bigger differences are to be expected due to the lower number of advertisements. The best performing variant on the test set is Variant 1.1. With an accuracy of 78.3% and a weighted average F_1-score of 0.783, it outperforms DeepL by 1.6% and 0.012 points respectively. The second-best model on the test set, Variant 1.2, is on par with DeepL in terms of accuracy and slightly below by -0.004 in terms of weighted average F_1-score.
Based on these results, the need for a larger test set is obvious. At the same time, it is highly interesting that different advertisements may cause performance differences to this degree.

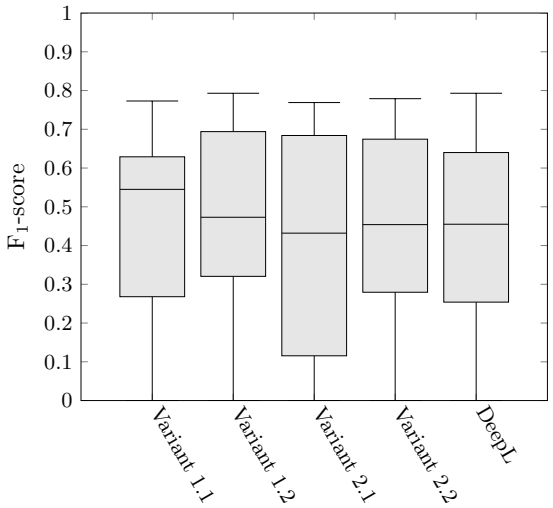

Figure 8.4: Unweighted F_1-scores on Variants using synthetic parallel data on the test set
English > German

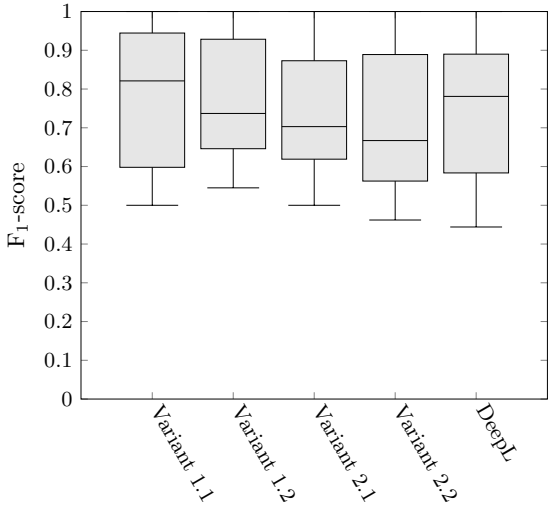

Figure 8.5: Unweighted F_1-scores on Variants using synthetic parallel data on the test set
Italian > German

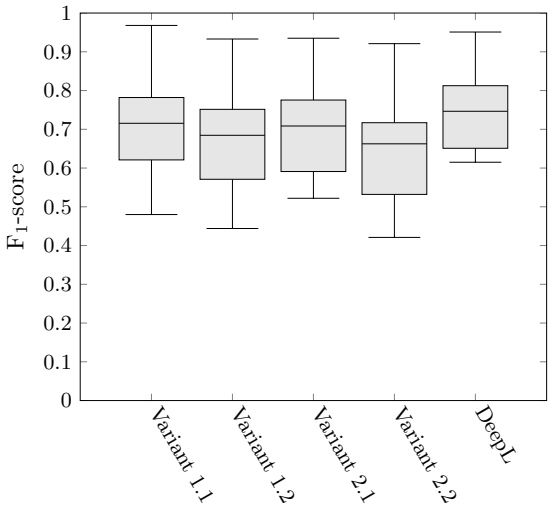

Figure 8.6: Unweighted F_1-scores on Variants using synthetic parallel data on the test set
French > German

9 Conclusion

In conclusion, it is possible to create a domain- and task-specific neural machine translation model that outperforms DeepL. Quality gains through different methods proposed by research could be repeated in this work. These included:

- Using multi-lingual models to boost performance on low resource languages, as well as providing domain adaption for languages lacking in-domain data.
- Quality gains through extending the context, which works even when it is introduced in domain-adaptation only.
- Using generic data to support domain-adaptation training.
- Using synthetic parallel data created by a first limited domain-adapted model:
  - In the form of back-translations.
  - And in the form of forward-translations combined with back-translations.

Whilst the efficacy of Elastic Weight Consolidation as proposed by Thompson et al. (2019) could be shown, in this work, with highly limited in-domain data, it was outperformed by domain adaption using generic data as support.

9.1 BLEU as a Performance Predictor

Based on the values of these eight variants and the generic base model on the development set, the correlation coefficient between BLEU-scores (against DeepL) and the overall accuracy on the English set is 0.733; between the BLEU-scores (against DeepL) and the weighted average F_1-scores the correlation coefficient is 0.765.
With regards to the French and Italian models, the correlation coefficients are much lower. The correlation coefficients between the BLEU-scores (against DeepL) and the overall accuracy is 0.445 and 0.468 for French and Italian respectively. Between the BLEU-scores (against DeepL) and the weighted average F_1-scores, the correlation coefficients of French and Italian respectively are at 0.417 and 0.472.
This difference could be due to performance differences of DeepL's models of these languages. Regardless of its cause, it once again serves as a caution against over-reliance on BLEU-scores.

9.2 Limitations

For this work, there were two main limitations: time and in-domain data. Limited time lead to earlier stopping of model training, especially of the generic base model, and no room for hyper-parameter tuning. The limited in-domain data especially held back training with Elastic Weight Consolidation, which may have caused it to underperform. In general, it can be assumed that

more human translations in their original state would have allowed for even higher performance. On the one hand, the use of these translations would not carry the risk of e.g. learning erroneous capitalisation. On the other hand, more in-domain data could improve the first domain-adapted, here called the alpha variants, models' performance. This higher performance, in turn, could improve the quality of the synthetic parallel data, increasing the benefit thereof even more. Alternatively, given more time, the quality of the synthetic in-domain data could be improved via iterative back-translation (Hoang et al., 2018).
An additional, but lesser, limitation was the implementation of EWC in Sockeye. At the time of writing this work, the version available merely runs on CPU. Through a small change of the code, it can be run on a single GPU. This limitation was exaggerated by the fact that using the empirical Fisher Information for each parameter requires ca. double the RAM. Thus, it was also necessary to halve the batch size. This caused training times to increase massively and may also have had an effect on quality.

9.3 Next Steps

Should SJMM and the Institute of Sociology at the University of Zurich wish to use an NMT-Model such as the one trained in this work, I present two recommendations.
Firstly, it is advisable to continue the training of the generic base model until convergence in order to maximise its performance. The inclusion of additional high-quality corpora may be of interest as well. This would require a new training start.
Secondly, I recommend to starting a project with the aim of creating additional human translations of advertisements, especially in Italian and French. This could be achieved by e.g. cooperating with a university of applied sciences which has an Institute of Translation. Once additional human translations are available, a retraining analogous to the best performing path of this work would be in order.

Independent of that, it could be interesting to have a closer look at the development and test sets, in order to try to determine the cause of the performance differences described in section 8.4.1.2.

Bibliography

(2021a). *Schweizer Berufsnomenklatur CH-ISCO-19*. Number 18004518.

(2021b). *Zuweisungsschlüssel der Schweizer Berufsnomenklatur CH-ISCO-19*. Number 18004516.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Bogoychev, N. and Sennrich, R. (2019). Domain, translationese and noise in synthetic data for neural machine translation. *CoRR*, abs/1911.03362.

Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Gnehm, A.-S. and Clematide, S. (2020). Text zoning and classification for job advertisements in German, French and English. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 83–93, Online. Association for Computational Linguistics.

Hieber, F., Domhan, T., Denkowski, M., and Vilar, D. (2020). Sockeye 2: A toolkit for neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 457–458, Lisboa, Portugal. European Association for Machine Translation.

Hoang, V. C. D., Koehn, P., Haffari, G., and Cohn, T. (2018). Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Hutchins, J., editor, *The Tenth Machine Translation Summit Proceedings of Conference*, pages 79–86. International Association for Machine Translation.

Lopes, A. V., Amin Farajian, M., Bawden, R., Zhang, M., and Martins, A. F. T. (2020). Document-level Neural MT: A Systematic Comparison. In *22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Post, M. and Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for 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 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Thompson, B., Gwinnup, J., Khayrallah, H., Duh, K., and Koehn, P. (2019). Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Vanmassenhove, E., Shterionov, D., and Gwilliam, M. (2021). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

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.

Appendix

Tables: Raw Human Translation

French>German: Raw Human Translation				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.500	0.800	0.615	5
Technology & Science	0.500	1.000	0.667	2
Trade & Sales	0.333	0.500	0.400	2
Office & Administration	0.500	0.200	0.286	5
Financial & Fiduciary Services	0.500	0.333	0.400	3
Hospitality & Personal Services	0.500	0.250	0.333	4
Health	1.000	1.000	1.000	1
Teaching & Public Services	0.750	0.750	0.750	8
Accuracy			**0.567**	30
Macro Average	0.573	0.604	**0.556**	30
Weighted Average	0.572	0.567	0.539	30

Table 9.1: Classification results of French>German translation on Raw Human Translation

English>German: Raw Human Translation				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.062	0.111	*0.080*	9
Technology & Science	0.426	0.769	0.548	26
IT	1.000	0.221	0.362	104
Trade & Sales	0.373	0.731	0.494	26
Office & Administration	0.395	0.484	0.435	31
Financial & Fiduciary Services	0.472	0.769	0.585	65
Management & Organisation	0.263	0.064	*0.103*	78
Hospitality & Personal Services	0.500	0.500	0.500	2
Health	0.154	0.286	*0.200*	7
Teaching & Public Services	0.245	0.650	0.356	20
Accuracy			0.405	368
Macro Average	0.389	0.459	0.366	368
Weighted Average	0.532	0.405	0.366	368

Table 9.2: Classification results of English>German translation on Raw Human Translation

Tables: DeepL on human-translated Set

French>German: Original DeepL-translation				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.500	0.600	0.545	5
Technology & Science	0.333	1.000	0.500	2
Trade & Sales	0.500	0.500	0.500	2
Office & Administration	0.667	0.400	0.500	5
Financial & Fiduciary Services	0.000	0.000	0.000	3
Hospitality & Personal Services	0.500	0.500	**0.500**	4
Health	1.000	1.000	1.000	1
Teaching & Public Services	1.000	0.625	**0.769**	8
Accuracy			0.533	30
Macro Average	0.500	0.514	0.479	30
Weighted Average	0.617	0.533	**0.546**	30

Table 9.3: Classification results of French>German translation on DeepL-translation

English>German: Original DeepL-translation				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.100	0.111	*0.105*	9
Technology & Science	0.372	0.615	0.464	26
IT	**0.872**	**0.721**	**0.789**	104
Trade & Sales	0.538	0.538	0.538	26
Office & Administration	0.459	0.548	0.500	31
Financial & Fiduciary Services	0.635	0.831	0.720	65
Management & Organisation	0.579	0.282	0.379	78
Hospitality & Personal Services	0.667	1.000	0.800	2
Health	0.500	0.429	0.462	7
Teaching & Public Services	0.382	0.650	0.481	20
Accuracy			**0.590**	368
Macro Average	0.511	0.573	**0.524**	368
Weighted Average	0.621	0.590	**0.585**	368

Table 9.4: Classification results of English>German translation on Raw Human Translation

Tables: English>German DeepL-translation Processing Experiments

English>German: Original DeepL-translation				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.1	0.111	0.105	9
Technology & Science	0.366	0.6	0.455	25
IT	0.88	0.717	0.79	92
Trade & Sales	0.52	0.52	0.52	25
Office & Administration	0.484	0.556	0.517	27
Financial & Fiduciary Services	0.65	0.825	0.727	63
Management & Organisation	0.595	0.306	0.404	72
Hospitality & Personal Services	0.667	1	0.8	2
Health	0.5	0.429	0.462	7
Teaching & Public Services	0.364	0.632	0.462	19
accuracy			0.589	341
macro average	0.512	0.57	0.524	341
weighted average	0.623	0.589	0.588	341

Table 9.5: Classification results of English>German translation on Original DeepL-translations (w/o multilingual advertisements).

English>German: DeepL-translation, Masked E-Mail & URL				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.091	0.111	0.1	9
Technology & Science	0.366	0.6	0.455	25
IT	0.865	0.696	0.771	92
Trade & Sales	0.56	0.56	0.56	25
Office & Administration	0.469	0.556	0.508	27
Financial & Fiduciary Services	0.646	0.81	0.718	63
Management & Organisation	0.568	0.292	0.385	72
Hospitality & Personal Services	0.667	1	0.8	2
Health	0.5	0.429	0.462	7
Teaching & Public Services	0.364	0.632	0.462	19
accuracy			0.581	341
macro average	0.509	0.568	0.522	341
weighted average	0.614	0.581	0.579	341

Table 9.6: Classification results of English>German translation on DeepL-translations (w/o multilingual advertisements). Masked e-mail addresses and URLs

English>German: DeepL-translation, Umlauts replaced				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0	0	0	9
Technology & Science	0.349	0.6	0.441	25
IT	0.851	0.685	0.759	92
Trade & Sales	0.577	0.6	0.588	25
Office & Administration	0.469	0.556	0.508	27
Financial & Fiduciary Services	0.63	0.81	0.708	63
Management & Organisation	0.571	0.278	0.374	72
Hospitality & Personal Services	0.667	1	0.8	2
Health	0.333	0.143	0.2	7
Teaching & Public Services	0.355	0.579	0.44	19
accuracy			0.566	341
macro average	0.48	0.525	0.482	341
weighted average	0.602	0.566	0.564	341

Table 9.7: Classification results of English>German translation on DeepL-translations (w/o multilingual advertisements). Umlauts replaced.

English>German: DeepL-translation, Digits replaced				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.091	0.111	0.1	9
Technology & Science	0.357	0.6	0.448	25
IT	0.833	0.707	0.765	92
Trade & Sales	0.519	0.56	0.538	25
Office & Administration	0.469	0.556	0.508	27
Financial & Fiduciary Services	0.641	0.794	0.709	63
Management & Organisation	0.594	0.264	0.365	72
Hospitality & Personal Services	0.667	1	0.8	2
Health	0.5	0.429	0.462	7
Teaching & Public Services	0.375	0.632	0.471	19
accuracy			0.575	341
macro average	0.505	0.565	0.517	341
weighted average	0.607	0.575	0.57	341

Table 9.8: Classification results of English>German translation on DeepL-translations (w/o multilingual advertisements). Digits replaced by 0.

English>German: DeepL-translation, Combined				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0	0	0	9
Technology & Science	0.385	0.6	0.469	25
IT	0.849	0.674	0.752	92
Trade & Sales	0.615	0.64	0.627	25
Office & Administration	0.438	0.519	0.475	27
Financial & Fiduciary Services	0.63	0.81	0.708	63
Management & Organisation	0.568	0.292	0.385	72
Hospitality & Personal Services	0.667	1	0.8	2
Health	0.4	0.286	0.333	7
Teaching & Public Services	0.375	0.632	0.471	19
accuracy			0.572	341
macro average	0.493	0.545	0.502	341
weighted average	0.606	0.572	0.571	341

Table 9.9: Classification results of English>German translation on DeepL-translations (w/o multilingual advertisements). Combined, true case.

English>German: DeepL-translation, Combined & lower-cased				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0	0	0	9
Technology & Science	0.475	0.76	0.585	25
IT	0.913	0.228	0.365	92
Trade & Sales	0.395	0.6	0.476	25
Office & Administration	0.387	0.444	0.414	27
Financial & Fiduciary Services	0.477	0.81	0.6	63
Management & Organisation	0.5	0.194	0.28	72
Hospitality & Personal Services	1	0.5	0.667	2
Health	0.25	0.286	0.267	7
Teaching & Public Services	0.167	0.474	0.247	19
accuracy			0.422	341
macro average	0.456	0.43	0.39	341
weighted average	0.555	0.422	0.402	341

Table 9.10: Classification results of English>German translation on DeepL-translations (w/o multilingual advertisements). Combined, lower-case.

Tables: French>German DeepL-translation Processing Experiments

French>German: Original DeepL-translation				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.6	0.6	0.6	5
Technology & Science	0.333	1	0.5	2
Trade & Sales	0.5	1	0.667	1
Office & Administration	0.667	0.4	0.5	5
Financial & Fiduciary Services	0	0	0	3
Hospitality & Personal Services	0.5	0.5	0.5	4
Health	1	1	1	1
Teaching & Public Services	1	0.625	0.769	8
accuracy			0.552	29
macro average	0.511	0.569	0.504	29
weighted average	0.638	0.552	0.563	29

Table 9.11: Classification results of French>German translation on Original DeepL-translations (w/o multilingual advertisements).

French>German: DeepL-translation, Masked E-Mail & URL				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.6	0.6	0.6	5
Technology & Science	0.333	1	0.5	2
Trade & Sales	0.5	1	0.667	1
Office & Administration	0.667	0.4	0.5	5
Financial & Fiduciary Services	0	0	0	3
Hospitality & Personal Services	0.5	0.5	0.5	4
Health	1	1	1	1
Teaching & Public Services	1	0.625	0.769	8
accuracy			0.552	29
macro average	0.511	0.569	0.504	29
weighted average	0.638	0.552	0.563	29

Table 9.12: Classification results of French>German translation on DeepL-translations (w/o multilingual advertisements). Masked e-mail addresses and URLs

French>German: DeepL-translation, Umlauts replaced				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.8	0.8	0.8	5
Technology & Science	0.4	1	0.571	2
Trade & Sales	0.5	1	0.667	1
Office & Administration	0.667	0.4	0.5	5
Financial & Fiduciary Services	0	0	0	3
Hospitality & Personal Services	0.75	0.75	0.75	4
Health	1	1	1	1
Teaching & Public Services	1	0.625	0.769	8
accuracy			0.621	29
macro average	0.569	0.619	0.562	29
weighted average	0.711	0.621	0.637	29

Table 9.13: Classification results of French>German translation on DeepL-translations (w/o multilingual advertisements). Umlauts replaced.

French>German: DeepL-translation, Digits replaced				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.6	0.6	0.6	5
Technology & Science	0.333	1	0.5	2
Trade & Sales	0.5	1	0.667	1
Office & Administration	0.667	0.4	0.5	5
Financial & Fiduciary Services	0	0	0	3
Hospitality & Personal Services	0.5	0.5	0.5	4
Health	1	1	1	1
Teaching & Public Services	1	0.625	0.769	8
accuracy			0.552	29
macro average	0.511	0.569	0.504	29
weighted average	0.638	0.552	0.563	29

Table 9.14: Classification results of French>German translation on DeepL-translations (w/o multilingual advertisements). Digits replaced by 0.

French>German: DeepL-translation, Combined				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.8	0.8	0.8	5
Technology & Science	0.4	1	0.571	2
Trade & Sales	0.5	1	0.667	1
Office & Administration	0.667	0.4	0.5	5
Financial & Fiduciary Services	0	0	0	3
Hospitality & Personal Services	0.75	0.75	0.75	4
Health	1	1	1	1
Teaching & Public Services	1	0.75	0.857	8
accuracy			0.655	29
macro average	0.569	0.633	0.572	29
weighted average	0.711	0.655	0.661	29

Table 9.15: Classification results of French>German translation on DeepL-translations (w/o multilingual advertisements). Combined, true case.

French>German: DeepL-translation, Combined & lower-cased				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.6	0.6	0.6	5
Technology & Science	0	0	0	2
Trade & Sales	0.2	1	0.333	1
Office & Administration	0	0	0	5
Financial & Fiduciary Services	0.5	0.333	0.4	3
Hospitality & Personal Services	1	0.25	0.4	4
Health	1	1	1	1
Teaching & Public Services	0.8	1	0.889	8
accuracy			0.571	29
macro average	0.456	0.465	0.402	29
weighted average	0.555	0.517	0.491	29

Table 9.16: Classification results of French>German translation on DeepL-translations (w/o multilingual advertisements). Combined, lower-case.

Alpha Variant Results

Baseline

English>German: Baseline variant				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.1	0.077	0.087	13
Technology & Science	0.5	0.706	0.585	51
IT	0.844	0.585	0.691	65
Trade & Sales	0.458	0.688	0.55	16
Office & Administration	0.379	0.407	0.393	27
Financial & Fiduciary Services	0.56	0.764	0.646	55
Management & Organisation	0.593	0.205	0.305	78
Hospitality & Personal Services	1	0.667	0.8	3
Health	0.375	0.429	0.4	7
Teaching & Public Services	0.392	0.714	0.506	28
accuracy			0.525	343
macro average	0.52	0.524	0.496	343
weighted average	0.562	0.525	0.507	343

Table 9.17: Classification results of English>German translation of the baseline variant

French>German: Baseline variant				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.611	0.635	0.623	52
Construction	0	0	0	11
Technology & Science	0.438	0.583	0.5	24
IT	0.909	0.588	0.714	17
Trade & Sales	0.648	0.686	0.667	51
Office & Administration	0.5	0.354	0.415	48
Financial & Fiduciary Services	0.68	0.791	0.731	43
Management & Organisation	0.286	0.316	0.3	19
Hospitality & Personal Services	0.938	0.366	0.526	41
Health	0.864	0.704	0.776	27
Teaching & Public Services	0.43	0.85	0.571	40
accuracy			0.582	373
macro average	0.573	0.534	0.529	373
weighted average	0.612	0.582	0.571	373

Table 9.18: Classification results of French>German translation of the baseline variant

Italian>German: Baseline variant				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.667	0.364	0.471	11
Construction	0	0	0	3
Technology & Science	0.375	0.6	0.462	5
IT	1	1	1	1
Trade & Sales	0.429	0.75	0.545	8
Office & Administration	0.75	0.429	0.545	7
Financial & Fiduciary Services	0.6	0.6	0.6	5
Management & Organisation	0	0	0	4
Hospitality & Personal Services	0.667	0.5	0.571	8
Health	1	0.571	0.727	7
Teaching & Public Services	0.083	0.333	0.133	3
accuracy			0.468	62
macro average	0.506	0.468	0.46	62
weighted average	0.556	0.468	0.479	62

Table 9.19: Classification results of Italian>German translation of the baseline variant

Variant 1

English>German: Alpha Variant 1				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.167	0.154	0.16	13
Technology & Science	0.506	0.765	0.609	51
IT	0.833	0.615	0.708	65
Trade & Sales	0.524	0.688	0.595	16
Office & Administration	0.417	0.556	0.476	27
Financial & Fiduciary Services	0.645	0.727	0.684	55
Management & Organisation	0.7	0.269	0.389	78
Hospitality & Personal Services	1	0.667	0.8	3
Health	0.5	0.286	0.364	7
Teaching & Public Services	0.412	0.75	0.532	28
accuracy			0.563	343
macro average	0.57	0.548	0.532	343
weighted average	0.612	0.563	0.552	343

Table 9.20: Classification results of English>German translation of Alpha Variant 1

French>German: Alpha Variant 1				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.633	0.596	0.614	52
Construction	1	0.182	0.308	11
Technology & Science	0.406	0.542	0.464	24
IT	0.786	0.647	0.71	17
Trade & Sales	0.625	0.588	0.606	51
Office & Administration	0.75	0.447	0.56	48
Financial & Fiduciary Services	0.739	0.791	0.764	43
Management & Organisation	0.212	0.368	0.269	19
Hospitality & Personal Services	0.846	0.297	0.44	41
Health	0.815	0.815	0.815	27
Teaching & Public Services	0.413	0.795	0.544	40
accuracy			0.58	373
macro average	0.657	0.552	0.554	373
weighted average	0.652	0.58	0.581	373

Table 9.21: Classification results of French>German translation of Alpha Variant 1

Italian>German: Alpha Variant 1				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.583	0.636	0.609	11
Construction	0	0	0	3
Technology & Science	0.286	0.4	0.333	5
IT	1	1	1	1
Trade & Sales	0.545	0.75	0.632	8
Office & Administration	0.667	0.571	0.615	7
Financial & Fiduciary Services	1	0.6	0.75	5
Management & Organisation	0	0	0	4
Hospitality & Personal Services	0.833	0.625	0.714	8
Health	1	0.571	0.727	7
Teaching & Public Services	0.25	0.667	0.364	3
accuracy			0.548	62
macro average	0.56	0.529	0.522	62
weighted average	0.601	0.548	0.554	62

Table 9.22: Classification results of Italian>German translation of Alpha Variant 1

Variant 2

English>German: Alpha Variant 2				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.333	0.231	0.273	13
Technology & Science	0.487	0.745	0.589	51
IT	0.824	0.646	0.724	65
Trade & Sales	0.565	0.812	0.667	16
Office & Administration	0.448	0.481	0.464	27
Financial & Fiduciary Services	0.559	0.691	0.618	55
Management & Organisation	0.724	0.269	0.393	78
Hospitality & Personal Services	1	0.667	0.8	3
Health	0.429	0.429	0.429	7
Teaching & Public Services	0.435	0.741	0.548	27
accuracy			0.564	342
macro average	0.58	0.571	0.55	342
weighted average	0.611	0.564	0.552	342

Table 9.23: Classification results of English>German translation of Alpha Variant 2

French>German: Alpha Variant 2				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.571	0.692	0.626	52
Construction	1	0.182	0.308	11
Technology & Science	0.406	0.542	0.464	24
IT	0.786	0.647	0.71	17
Trade & Sales	0.611	0.647	0.629	51
Office & Administration	0.545	0.375	0.444	47
Financial & Fiduciary Services	0.745	0.884	0.809	43
Management & Organisation	0.217	0.263	0.238	19
Hospitality & Personal Services	0.938	0.366	0.526	37
Health	0.821	0.852	0.836	27
Teaching & Public Services	0.548	0.85	0.667	39
accuracy			0.611	367
macro average	0.66	0.573	0.573	367
weighted average	0.648	0.611	0.6	367

Table 9.24: Classification results of French>German translation of Alpha Variant 2

Italian>German: Alpha Variant 2				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.5	0.5	0.5	10
Construction	1	0.333	0.5	3
Technology & Science	0.2	0.2	0.2	5
IT	1	1	1	1
Trade & Sales	0.6	0.75	0.667	8
Office & Administration	0.556	0.714	0.625	7
Financial & Fiduciary Services	0.75	0.6	0.667	5
Management & Organisation	0	0	0	4
Hospitality & Personal Services	0.833	0.714	0.769	7
Health	1	0.571	0.727	7
Teaching & Public Services	0.429	1	0.6	3
accuracy			0.567	60
macro average	0.624	0.58	0.569	60
weighted average	0.609	0.567	0.564	60

Table 9.25: Classification results of Italian>German translation of Alpha Variant 2

Variant 3

English>German: Alpha Variant 3				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.286	0.154	0.2	13
Technology & Science	0.487	0.725	0.583	51
IT	0.885	0.708	0.786	65
Trade & Sales	0.48	0.75	0.585	16
Office & Administration	0.548	0.63	0.586	27
Financial & Fiduciary Services	0.641	0.745	0.689	55
Management & Organisation	0.75	0.308	0.436	78
Hospitality & Personal Services	1	0.667	0.8	3
Health	0.4	0.286	0.333	7
Teaching & Public Services	0.417	0.741	0.533	27
accuracy			0.594	342
macro average	0.589	0.571	0.553	342
weighted average	0.641	0.594	0.584	342

Table 9.26: Classification results of English>German translation of Alpha Variant 3

French>German: Alpha Variant 3				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.625	0.577	0.6	52
Construction	1	0.273	0.429	11
Technology & Science	0.378	0.583	0.459	24
IT	0.812	0.765	0.788	17
Trade & Sales	0.647	0.647	0.647	51
Office & Administration	0.844	0.574	0.684	47
Financial & Fiduciary Services	0.795	0.814	0.805	43
Management & Organisation	0.364	0.421	0.39	19
Hospitality & Personal Services	0.733	0.297	0.423	37
Health	0.786	0.815	0.8	27
Teaching & Public Services	0.451	0.821	0.582	39
accuracy			0.621	367
macro average	0.676	0.599	0.601	367
weighted average	0.671	0.621	0.62	367

Table 9.27: Classification results of French>German translation of Alpha Variant 3

Italian>German: Alpha Variant 3				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.667	0.6	0.632	10
Construction	0	0	0	3
Technology & Science	0.333	0.4	0.364	5
IT	1	1	1	1
Trade & Sales	0.778	0.875	0.824	8
Office & Administration	0.571	0.571	0.571	7
Financial & Fiduciary Services	1	0.8	0.889	5
Management & Organisation	0.286	0.5	0.364	4
Hospitality & Personal Services	0.833	0.714	0.769	7
Health	1	0.571	0.727	7
Teaching & Public Services	0.5	1	0.667	3
accuracy			0.633	60
macro average	0.633	0.639	0.619	60
weighted average	0.667	0.633	0.635	60

Table 9.28: Classification results of Italian>German translation of Alpha Variant 3

Variant 4

English>German: Alpha Variant 4				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.25	0.154	0.19	13
Technology & Science	0.507	0.745	0.603	51
IT	0.86	0.662	0.748	65
Trade & Sales	0.684	0.812	0.743	16
Office & Administration	0.4	0.519	0.452	27
Financial & Fiduciary Services	0.6	0.764	0.672	55
Management & Organisation	0.667	0.282	0.396	78
Hospitality & Personal Services	1	0.667	0.8	3
Health	0.333	0.286	0.308	7
Teaching & Public Services	0.455	0.741	0.563	27
accuracy			0.579	342
macro average	0.576	0.563	0.548	342
weighted average	0.612	0.579	0.566	342

Table 9.29: Classification results of English>German translation of Alpha Variant 4

French>German: Alpha Variant 4				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.564	0.596	0.579	52
Construction	0.6	0.273	0.375	11
Technology & Science	0.419	0.542	0.473	24
IT	0.846	0.647	0.733	17
Trade & Sales	0.688	0.647	0.667	51
Office & Administration	0.686	0.511	0.585	47
Financial & Fiduciary Services	0.756	0.791	0.773	43
Management & Organisation	0.3	0.474	0.367	19
Hospitality & Personal Services	0.8	0.324	0.462	37
Health	0.84	0.778	0.808	27
Teaching & Public Services	0.523	0.872	0.654	39
accuracy			0.613	367
macro average	0.638	0.587	0.589	367
weighted average	0.65	0.613	0.611	367

Table 9.30: Classification results of French>German translation of Alpha Variant 4

Italian>German: Alpha Variant 4				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.5	0.6	0.545	10
Construction	0	0	0	3
Technology & Science	0.4	0.4	0.4	5
IT	1	1	1	1
Trade & Sales	0.444	0.5	0.471	8
Office & Administration	0.8	0.571	0.667	7
Financial & Fiduciary Services	1	0.8	0.889	5
Management & Organisation	0.125	0.25	0.167	4
Hospitality & Personal Services	0.75	0.429	0.545	7
Health	1	0.714	0.833	7
Teaching & Public Services	0.429	1	0.6	3
accuracy			0.55	60
macro average	0.586	0.569	0.556	60
weighted average	0.603	0.55	0.557	60

Table 9.31: Classification results of Italian>German translation of Alpha Variant 4

Final Model Results - Development Set

Variant 1.1

English>German: Variant 1.1 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.385	0.385	0.385	13
Technology & Science	0.521	0.745	0.613	51
IT	0.907	0.754	0.824	65
Trade & Sales	0.48	0.75	0.585	16
Office & Administration	0.536	0.556	0.545	27
Financial & Fiduciary Services	0.651	0.745	0.695	55
Management & Organisation	0.75	0.308	0.436	78
Hospitality & Personal Services	1	0.667	0.8	3
Health	0.6	0.429	0.5	7
Teaching & Public Services	0.404	0.704	0.514	27
accuracy			0.608	342
macro average	0.623	0.604	0.59	342
weighted average	0.658	0.608	0.602	342

Table 9.32: Classification results of English>German translation of Variant 1.1 on the development set

French>German: Variant 1.1 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.731	0.731	0.731	52
Construction	0.625	0.455	0.526	11
Technology & Science	0.353	0.5	0.414	24
IT	0.786	0.647	0.71	17
Trade & Sales	0.735	0.706	0.72	51
Office & Administration	0.625	0.426	0.506	47
Financial & Fiduciary Services	0.75	0.767	0.759	43
Management & Organisation	0.333	0.368	0.35	19
Hospitality & Personal Services	0.783	0.486	0.6	37
Health	0.76	0.704	0.731	27
Teaching & Public Services	0.492	0.821	0.615	39
accuracy			0.629	367
macro average	0.634	0.601	0.606	367
weighted average	0.656	0.629	0.631	367

Table 9.33: Classification results of French>German translation of Variant 1.1 on the development set

Italian>German: Variant 1.1 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.5	0.6	0.545	10
Construction	0	0	0	3
Technology & Science	0.4	0.4	0.4	5
IT	1	1	1	1
Trade & Sales	0.444	0.5	0.471	8
Office & Administration	0.8	0.571	0.667	7
Financial & Fiduciary Services	1	0.8	0.889	5
Management & Organisation	0.125	0.25	0.167	4
Hospitality & Personal Services	0.75	0.429	0.545	7
Health	1	0.714	0.833	7
Teaching & Public Services	0.429	1	0.6	3
accuracy			0.55	60
macro average	0.586	0.569	0.556	60
weighted average	0.603	0.55	0.557	60

Table 9.34: Classification results of Italian>German translation of Variant 1.1 on the development set

Variant 1.2

English>German: Variant 1.2 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.444	0.308	0.364	13
Technology & Science	0.487	0.725	0.583	51
IT	0.887	0.723	0.797	65
Trade & Sales	0.444	0.75	0.558	16
Office & Administration	0.485	0.593	0.533	27
Financial & Fiduciary Services	0.667	0.727	0.696	55
Management & Organisation	0.714	0.256	0.377	78
Hospitality & Personal Services	1	0.667	0.8	3
Health	0.5	0.286	0.364	7
Teaching & Public Services	0.4	0.741	0.519	27
accuracy			0.585	342
macro average	0.603	0.578	0.559	342
weighted average	0.638	0.585	0.574	342

Table 9.35: Classification results of English>German translation of Variant 1.2 on the development set

French>German: Variant 1.2 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.661	0.712	0.685	52
Construction	0.6	0.545	0.571	11
Technology & Science	0.484	0.625	0.545	24
IT	0.8	0.706	0.75	17
Trade & Sales	0.702	0.647	0.673	51
Office & Administration	0.706	0.511	0.593	47
Financial & Fiduciary Services	0.786	0.767	0.776	43
Management & Organisation	0.28	0.368	0.318	19
Hospitality & Personal Services	0.842	0.432	0.571	37
Health	0.767	0.852	0.807	27
Teaching & Public Services	0.586	0.872	0.701	39
accuracy			0.654	367
macro average	0.656	0.64	0.636	367
weighted average	0.678	0.654	0.653	367

Table 9.36: Classification results of French>German translation of Variant 1.2 on the development set

Italian>German: Variant 1.2 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.5	0.6	0.545	10
Construction	0.333	0.333	0.333	3
Technology & Science	0.286	0.4	0.333	5
IT	0.5	1	0.667	1
Trade & Sales	0.778	0.875	0.824	8
Office & Administration	0.6	0.429	0.5	7
Financial & Fiduciary Services	1	0.6	0.75	5
Management & Organisation	0.2	0.25	0.222	4
Hospitality & Personal Services	0.833	0.714	0.769	7
Health	1	0.571	0.727	7
Teaching & Public Services	0.5	0.667	0.571	3
accuracy			0.583	60
macro average	0.594	0.585	0.567	60
weighted average	0.641	0.583	0.595	60

Table 9.37: Classification results of Italian>German translation of Variant 1.2 on the development set

Variant 2.1

English>German: Variant 2.1 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.455	0.385	0.417	13
Technology & Science	0.479	0.686	0.565	51
IT	0.868	0.708	0.78	65
Trade & Sales	0.5	0.75	0.6	16
Office & Administration	0.5	0.593	0.542	27
Financial & Fiduciary Services	0.661	0.709	0.684	55
Management & Organisation	0.828	0.308	0.449	78
Hospitality & Personal Services	1	0.667	0.8	3
Health	0.333	0.429	0.375	7
Teaching & Public Services	0.42	0.778	0.545	27
accuracy			0.594	342
macro average	0.604	0.601	0.576	342
weighted average	0.66	0.594	0.589	342

Table 9.38: Classification results of English>German translation of Variant 2.1 on the development set

French>German: Variant 2.1 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.648	0.673	0.66	52
Construction	0.75	0.273	0.4	11
Technology & Science	0.438	0.583	0.5	24
IT	0.824	0.824	0.824	17
Trade & Sales	0.667	0.627	0.646	51
Office & Administration	0.667	0.468	0.55	47
Financial & Fiduciary Services	0.733	0.767	0.75	43
Management & Organisation	0.346	0.474	0.4	19
Hospitality & Personal Services	0.833	0.405	0.545	37
Health	0.786	0.815	0.8	27
Teaching & Public Services	0.548	0.872	0.673	39
accuracy			0.635	367
macro average	0.658	0.616	0.614	367
weighted average	0.663	0.635	0.631	367

Table 9.39: Classification results of French>German translation of Variant 2.1 on the development set

Italian>German: Variant 2.1 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.5	0.6	0.545	10
Construction	1	0.333	0.5	3
Technology & Science	0.429	0.6	0.5	5
IT	0.5	1	0.667	1
Trade & Sales	0.75	0.75	0.75	8
Office & Administration	0.5	0.429	0.462	7
Financial & Fiduciary Services	0.75	0.6	0.667	5
Management & Organisation	0.333	0.25	0.286	4
Hospitality & Personal Services	1	0.714	0.833	7
Health	1	0.571	0.727	7
Teaching & Public Services	0.375	1	0.545	3
accuracy			0.6	60
macro average	0.649	0.623	0.589	60
weighted average	0.673	0.6	0.606	60

Table 9.40: Classification results of Italian>German translation of Variant 2.1 on the development set

Variant 2.2

English>German: Variant 2.2 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.429	0.231	0.3	13
Technology & Science	0.442	0.667	0.531	51
IT	0.855	0.723	0.783	65
Trade & Sales	0.458	0.688	0.55	16
Office & Administration	0.485	0.593	0.533	27
Financial & Fiduciary Services	0.656	0.727	0.69	55
Management & Organisation	0.724	0.269	0.393	78
Hospitality & Personal Services	1	0.667	0.8	3
Health	0.5	0.286	0.364	7
Teaching & Public Services	0.4	0.741	0.519	27
accuracy			0.573	342
macro average	0.595	0.559	0.546	342
weighted average	0.625	0.573	0.563	342

Table 9.41: Classification results of English>German translation of Variant 2.2 on the development set

French>German: Variant 2.2 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.692	0.692	0.692	52
Construction	0.6	0.273	0.375	11
Technology & Science	0.469	0.625	0.536	24
IT	0.786	0.647	0.71	17
Trade & Sales	0.66	0.608	0.633	51
Office & Administration	0.71	0.468	0.564	47
Financial & Fiduciary Services	0.72	0.837	0.774	43
Management & Organisation	0.36	0.474	0.409	19
Hospitality & Personal Services	0.8	0.432	0.561	37
Health	0.821	0.852	0.836	27
Teaching & Public Services	0.524	0.846	0.647	39
accuracy			0.64	367
macro average	0.649	0.614	0.613	367
weighted average	0.665	0.64	0.636	367

Table 9.42: Classification results of French>German translation of Variant 2.2 on the development set

Italian>German: Variant 2.2 on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.7	0.7	0.7	10
Construction	1	0.333	0.5	3
Technology & Science	0.333	0.4	0.364	5
IT	0.5	1	0.667	1
Trade & Sales	0.7	0.875	0.778	8
Office & Administration	0.714	0.714	0.714	7
Financial & Fiduciary Services	1	0.6	0.75	5
Management & Organisation	0	0	0	4
Hospitality & Personal Services	0.857	0.857	0.857	7
Health	1	0.571	0.727	7
Teaching & Public Services	0.429	1	0.6	3
accuracy			0.65	60
macro average	0.658	0.641	0.605	60
weighted average	0.701	0.65	0.647	60

Table 9.43: Classification results of Italian>German translation of Variant 2.2 on the development set

EWC Final

English>German: EWC on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.222	0.154	0.182	13
Technology & Science	0.471	0.784	0.588	51
IT	0.837	0.631	0.719	65
Trade & Sales	0.462	0.75	0.571	16
Office & Administration	0.433	0.481	0.456	27
Financial & Fiduciary Services	0.582	0.709	0.639	55
Management & Organisation	0.607	0.218	0.321	78
Hospitality & Personal Services	1	0.667	0.8	3
Health	0.5	0.571	0.533	7
Teaching & Public Services	0.474	0.667	0.554	27
accuracy			0.55	342
macro average	0.559	0.563	0.536	342
weighted average	0.582	0.55	0.532	342

Table 9.44: Classification results of English>German translation of the EWC model on the development set

French>German: EWC on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.587	0.712	0.643	52
Construction	0.714	0.455	0.556	11
Technology & Science	0.387	0.5	0.436	24
IT	0.786	0.647	0.71	17
Trade & Sales	0.667	0.627	0.646	51
Office & Administration	0.703	0.553	0.619	47
Financial & Fiduciary Services	0.833	0.814	0.824	43
Management & Organisation	0.15	0.158	0.154	19
Hospitality & Personal Services	0.826	0.514	0.633	37
Health	0.76	0.704	0.731	27
Teaching & Public Services	0.579	0.846	0.688	39
accuracy			0.632	367
macro average	0.636	0.594	0.604	367
weighted average	0.655	0.632	0.633	367

Table 9.45: Classification results of French>German translation of the EWC model on the development set

Italian>German: EWC on the Dev Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.625	0.5	0.556	10
Construction	0.333	0.333	0.333	3
Technology & Science	0.333	0.4	0.364	5
IT	1	1	1	1
Trade & Sales	0.6	0.75	0.667	8
Office & Administration	0.714	0.714	0.714	7
Financial & Fiduciary Services	0.6	0.6	0.6	5
Management & Organisation	0	0	0	4
Hospitality & Personal Services	0.714	0.714	0.714	7
Health	1	0.571	0.727	7
Teaching & Public Services	0.375	1	0.545	3
accuracy			0.65	60
macro average	0.658	0.641	0.605	60
weighted average	0.701	0.65	0.647	60

Table 9.46: Classification results of Italian>German translation of the EWC model on the development set

Final Model Results - Test Set

DeepL-baseline

English>German: DeepL-baseline on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.25	0.4	0.308	5
Construction	0	0	0	1
Technology & Science	0.511	0.818	0.629	55
IT	0.868	0.73	0.793	63
Trade & Sales	0.667	0.636	0.651	22
Office & Administration	0.5	0.56	0.528	25
Financial & Fiduciary Services	0.625	0.804	0.703	56
Management & Organisation	0.553	0.292	0.382	72
Hospitality & Personal Services	0.5	0.25	0.333	4
Health	0.333	0.143	0.2	14
Teaching & Public Services	0.538	0.56	0.549	25
accuracy			0.596	342
macro average	0.486	0.472	0.462	342
weighted average	0.603	0.596	0.58	342

Table 9.47: Classification results of English>German translation of DeepL on the test set

French>German: DeepL-baseline on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.701	0.825	0.758	57
Construction	0.875	0.538	0.667	13
Technology & Science	0.718	0.737	0.727	38
IT	0.923	0.75	0.828	16
Trade & Sales	0.83	0.796	0.812	49
Office & Administration	0.714	0.571	0.635	35
Financial & Fiduciary Services	0.743	0.788	0.765	33
Management & Organisation	0.606	0.625	0.615	32
Hospitality & Personal Services	0.857	0.643	0.735	28
Health	0.935	0.967	0.951	30
Teaching & Public Services	0.74	0.902	0.813	41
accuracy			0.632	367
macro average	0.636	0.594	0.604	367
weighted average	0.655	0.632	0.633	367

Table 9.48: Classification results of French>German translation of DeepL on the test set

Italian>German: DeepL-baseline on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	1	0.7	0.824	10
Construction	0.5	1	0.667	1
Technology & Science	0.4	0.5	0.444	4
IT	0.667	1	0.8	2
Trade & Sales	0.8	0.727	0.762	11
Office & Administration	0.5	0.5	0.5	6
Financial & Fiduciary Services	1	1	1	4
Management & Organisation	1	0.75	0.857	4
Hospitality & Personal Services	0.667	0.75	0.706	8
Health	1	1	1	4
Teaching & Public Services	0.857	1	0.923	6
accuracy			0.767	60
macro average	0.763	0.812	0.771	60
weighted average 0.795	0.767	0.771	60	

Table 9.49: Classification results of Italian>German translation of DeepL on the test set

Variant 1.1

English>German: Variant 1.1 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.222	0.4	0.286	5
Construction	0	0	0	1
Technology & Science	0.545	0.764	0.636	55
IT	0.821	0.73	0.773	63
Trade & Sales	0.609	0.636	0.622	22
Office & Administration	0.483	0.56	0.519	25
Financial & Fiduciary Services	0.597	0.821	0.692	56
Management & Organisation	0.607	0.236	0.34	72
Hospitality & Personal Services	0.667	0.5	0.571	4
Health	0.333	0.2	0.25	15
Teaching & Public Services	0.515	0.654	0.576	25
accuracy			0.59	344
macro average	0.491	0.5	0.479	344
weighted average	0.6	0.59	0.57	344

Table 9.50: Classification results of English>German translation of Variant 1.1 on the test set

French>German: Variant 1.1 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.692	0.789	0.738	57
Construction	0.5	0.462	0.48	13
Technology & Science	0.667	0.737	0.7	38
IT	0.9	0.562	0.692	16
Trade & Sales	0.848	0.796	0.821	49
Office & Administration	0.743	0.743	0.743	35
Financial & Fiduciary Services	0.867	0.788	0.825	33
Management & Organisation	0.7	0.656	0.677	32
Hospitality & Personal Services	0.722	0.464	0.565	28
Health	0.938	1	0.968	30
Teaching & Public Services	0.654	0.829	0.731	41
accuracy			0.745	372
macro average	0.748	0.712	0.722	372
weighted average	0.751	0.745	0.742	372

Table 9.51: Classification results of French>German translation of Variant 1.1 on the test set

Italian>German: Variant 1.1 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.889	0.8	0.842	10
Construction	0.333	1	0.5	1
Technology & Science	1	0.75	0.857	4
IT	1	1	1	2
Trade & Sales	0.667	0.727	0.696	11
Office & Administration	0.5	0.5	0.5	6
Financial & Fiduciary Services	1	1	1	4
Management & Organisation	1	1	1	4
Hospitality & Personal Services	0.857	0.75	0.8	8
Health	0.8	1	0.889	4
Teaching & Public Services	0.8	0.667	0.727	6
accuracy			0.783	60
macro average	0.804	0.836	0.801	60
weighted average	0.807	0.783	0.789	60

Table 9.52: Classification results of Italian>German translation of Variant 1.1 on the test set

Variant 1.2

English>German: Variant 1.2 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.286	0.4	0.333	5
Construction	0	0	0	1
Technology & Science	0.581	0.782	0.667	55
IT	0.833	0.714	0.769	63
Trade & Sales	0.762	0.727	0.744	22
Office & Administration	0.567	0.68	0.618	25
Financial & Fiduciary Services	0.603	0.839	0.701	56
Management & Organisation	0.645	0.278	0.388	72
Hospitality & Personal Services	0	0	0	4
Health	0.273	0.2	0.231	15
Teaching & Public Services	0.405	0.577	0.476	26
accuracy			0.605	344
macro average	0.45	0.472	0.448	344
weighted average	0.615	0.605	0.586	344

Table 9.53: Classification results of English>German translation of Variant 1.2 on the test set

French>German: Variant 1.2 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.698	0.772	0.733	57
Construction	0.6	0.462	0.522	13
Technology & Science	0.689	0.816	0.747	38
IT	0.9	0.562	0.692	16
Trade & Sales	0.86	0.755	0.804	49
Office & Administration	0.676	0.657	0.667	35
Financial & Fiduciary Services	0.867	0.788	0.825	33
Management & Organisation	0.585	0.75	0.658	32
Hospitality & Personal Services	0.786	0.393	0.524	28
Health	0.906	0.967	0.935	30
Teaching & Public Services	0.66	0.805	0.725	41
accuracy			0.734	372
macro average	0.748	0.702	0.712	372
weighted average	0.746	0.734	0.73	372

Table 9.54: Classification results of French>German translation of Variant 1.2 on the test set

Italian>German: Variant 1.2 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.875	0.7	0.778	10
Construction	0.333	1	0.5	1
Technology & Science	1	0.75	0.857	4
IT	1	0.5	0.667	2
Trade & Sales	0.778	0.636	0.7	11
Office & Administration	0.667	0.667	0.667	6
Financial & Fiduciary Services	1	1	1	4
Management & Organisation	1	1	1	4
Hospitality & Personal Services	0.667	0.75	0.706	8
Health	0.8	1	0.889	4
Teaching & Public Services	0.5	0.667	0.571	6
accuracy			0.75	60
macro average	0.784	0.788	0.758	60
weighted average	0.786	0.75	0.756	60

Table 9.55: Classification results of Italian>German translation of Variant 1.2 on the test set

Variant 2.1

English>German: Variant 2.1 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.4	0.4	0.4	5
Construction	0	0	0	1
Technology & Science	0.587	0.8	0.677	55
IT	0.868	0.73	0.793	63
Trade & Sales	0.696	0.727	0.711	22
Office & Administration	0.536	0.6	0.566	25
Financial & Fiduciary Services	0.636	0.875	0.737	56
Management & Organisation	0.595	0.306	0.404	72
Hospitality & Personal Services	0.5	0.25	0.333	4
Health	0.364	0.267	0.308	15
Teaching & Public Services	0.485	0.615	0.542	26
accuracy			0.625	344
macro average	0.515	0.506	0.497	344
weighted average	0.628	0.625	0.609	344

Table 9.56: Classification results of English>German translation of Variant 2.1 on the test set

French>German: Variant 2.1 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.662	0.754	0.705	57
Construction	0.8	0.308	0.444	13
Technology & Science	0.667	0.684	0.675	38
IT	0.909	0.625	0.741	16
Trade & Sales	0.804	0.755	0.779	49
Office & Administration	00.676	0.714	0.694	35
Financial & Fiduciary Services	0.8	0.727	0.762	33
Management & Organisation	0.514	0.594	0.551	32
Hospitality & Personal Services	0.812	0.464	0.591	28
Health	0.933	0.933	0.933	30
Teaching & Public Services	0.571	0.78	0.66	41
accuracy			0.702	372
macro average	0.741	0.667	0.685	372
weighted average	0.721	0.702	0.7	372

Table 9.57: Classification results of French>German translation of Variant 2.1 on the test set

Italian>German: Variant 2.1 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.875	0.7	0.778	10
Construction	0.5	1	0.667	1
Technology & Science	1	0.75	0.857	4
IT	1	0.5	0.667	2
Trade & Sales	0.667	0.727	0.696	11
Office & Administration	0.6	0.5	0.545	6
Financial & Fiduciary Services	1	1	1	4
Management & Organisation	1	1	1	4
Hospitality & Personal Services	0.625	0.625	0.625	8
Health	1	1	1	4
Teaching & Public Services	0.667	1	0.8	6
accuracy			0.767	60
macro average	0.812	0.8	0.785	60
weighted average	0.786	0.767	0.766	60

Table 9.58: Classification results of Italian>German translation of Variant 2.1 on the test set

Variant 2.2

English>German: Variant 2.2 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.222	0.4	0.286	5
Construction	0	0	0	1
Technology & Science	0.581	0.782	0.667	55
IT	0.88	0.698	0.779	63
Trade & Sales	0.682	0.682	0.682	22
Office & Administration	0.516	0.64	0.571	25
Financial & Fiduciary Services	0.622	0.821	0.708	56
Management & Organisation	0.579	0.306	0.4	72
Hospitality & Personal Services	0.5	0.25	0.333	4
Health	0.429	0.2	0.273	15
Teaching & Public Services	0.432	0.615	0.508	25
accuracy			0.605	344
macro average	0.495	0.49	0.473	344
weighted average	0.618	0.605	0.592	344

Table 9.59: Classification results of English>German translation of Variant 2.2 on the test set

French>German: Variant 2.2 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.667	0.772	0.715	57
Construction	0.667	0.308	0.421	13
Technology & Science	0.658	0.658	0.658	38
IT	0.9	0.562	0.692	16
Trade & Sales	0.755	0.755	0.755	49
Office & Administration	0.568	0.6	0.583	35
Financial & Fiduciary Services	0.742	0.697	0.719	33
Management & Organisation	0.559	0.594	0.576	32
Hospitality & Personal Services	0.769	0.357	0.488	28
Health	0.879	0.967	0.921	30
Teaching & Public Services	0.582	0.78	0.667	41
accuracy			0.68	372
macro average	0.704	0.641	0.654	372
weighted average	0.691	0.68	0.673	372

Table 9.60: Classification results of French>German translation of Variant 2.2 on the test set

Italian>German: Variant 2.2 on the Test Set				
Label	Precision	Recall	F_1-score	Support
Industry & Transport	0.875	0.7	0.778	10
Construction	0.333	1	0.5	1
Technology & Science	1	0.5	0.667	4
IT	1	0.5	0.667	2
Trade & Sales	0.636	0.636	0.636	11
Office & Administration	0.429	0.5	0.462	6
Financial & Fiduciary Services	1	1	1	4
Management & Organisation	1	1	1	4
Hospitality & Personal Services	0.625	0.625	0.625	8
Health	1	1	1	4
Teaching & Public Services	0.625	0.833	0.714	6
accuracy			0.717	60
macro average	0.775	0.754	0.732	60
weighted average	0.757	0.717	0.722	60

Table 9.61: Classification results of Italian>German translation of Variant 2.2 on the test set

Hyper Parameters

Baseline Model

allow_missing_params: false
amp: false
amp_scale_interval: 2000
batch_sentences_multiple_of: 8
batch_size: 4096
batch_type: word
bucket_scaling: false
bucket_width: 8
cache_last_best_params: 0
cache_metric: perplexity
cache_strategy: best
checkpoint_improvement_threshold: 0.0
checkpoint_interval: 5000
config: null
decode_and_evaluate: 3000
decoder: transformer
dtype: float32
encoder: transformer
env: null
gradient_clipping_threshold: 1.0
gradient_clipping_type: none
ignore_extra_params: false
initial_learning_rate: 0.0002
keep_initializations: false
keep_last_params: -1
kvstore: device
label_smoothing: 0.1
learning_rate_reduce_factor: 0.9
learning_rate_reduce_num_not_improved: 8
learning_rate_scheduler_type: plateau-reduce
learning_rate_t_scale: 1.0
learning_rate_warmup: 0
length_task: null
length_task_layers: 1
length_task_weight: 1.0
lock_dir: /tmp
loss: cross-entropy-without-softmax-output
max_checkpoints: null
max_num_checkpoint_not_improved: 8
max_num_epochs: null
max_samples: null
max_seconds: null
max_seq_len:
- 95

- 95
max_updates: 500000
min_num_epochs: null
min_samples: null
min_updates: null
momentum: null
monitor_pattern: null
monitor_stat_func: mx_default
no_bucket_scaling: null
no_bucketing: false
no_hybridization: false
no_logfile: false
num_embed:
- null
- null
num_layers:
- 6
- 6
num_words:
- 0
- 0
optimized_metric: perplexity
optimizer: adam
optimizer_params: null
overwrite_output: false
pad_vocab_to_multiple_of: null
params: null
quiet: false
quiet_secondary_workers: false
round_batch_sizes_to_multiple_of: null
seed: 1
transformer_activation_type:
- relu
- relu
transformer_attention_heads:
- 8
- 8
transformer_dropout_act: &id001
- 0.1
- 0.1
transformer_dropout_attention: *id001
transformer_dropout_prepost: *id001
transformer_feed_forward_num_hidden:
- 2048
- 2048
transformer_feed_forward_use_glu: false
transformer_model_size:
- 512

- 512
transformer_positional_embedding_type: fixed
transformer_postprocess:
- dr
- dr
transformer_preprocess:
- n
- n
update_interval: 1
weight_decay: 0.0
weight_init: xavier
weight_init_scale: 3.0
weight_init_xavier_factor_type: avg
weight_init_xavier_rand_type: uniform
weight_tying_type: src_trg_softmax
word_min_count:
- 1
- 1