Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3038–3047
Marseille, 20-25 June 2022

Unsupervised Machine Translation in Real-World Scenarios

Ona de Gibert[1], Iakes Goenaga[2], Jordi Armengol-Estapé[1], Olatz Perez-de-Viñaspre[2]
Carla Parra[3], Marina Sánchez-Torrón[4], Marcis Pinnis[5], Gorka Labaka[2], Maite Melero[1]
Barcelona Supercomputing Center, Barcelona, Spain[1]
HiTZ zentroa - Ixa, Euskal Herriko Unibertsitatea UPV/EHU, Donostia, Spain[2]
RWS Language Weaver, Dublin, Ireland[3]
Unbabel, Lisbon, Portugal[4]
Tilde, Riga, Latvia[5]
{ona.degibert, jordi.armengol, maite.melero}@bsc.es[1]
{iakes.goenaga, olatz.perezdevinaspre, gorka.labaka}@ehu.eus[2]
CParraEscartin@rws.com[3], marina.sanchez@unbabel.com[4], marcis.pinnis@tilde.lv[5]

Abstract

In this work, we present the work that has been carried on in the MT4All CEF project and the resources that it has generated by leveraging recent research carried out in the field of unsupervised learning. In the course of the project 18 monolingual corpora for specific domains and languages have been collected, and 12 bilingual dictionaries and translation models have been generated. As part of the research, the unsupervised MT methodology based only on monolingual corpora (Artetxe et al., 2017) has been tested on a variety of languages and domains. Results show that in specialised domains, when there is enough monolingual in-domain data, unsupervised results are comparable to those of general domain supervised translation, and that, at any rate, unsupervised techniques can be used to boost results whenever very little data is available.

Keywords: Machine Translation, Unsupervised, Neural Networks, Low-resource languages

1. Introduction

The foundation of Machine Translation (MT) largely depends on the availability of large parallel corpus of the targeted language pair. Nevertheless, large amounts of bilingual corpora are not always available, specially for specific domains or low-resource languages. In recent years, to address this issue, unsupervised techniques have appeared which rely solely or partially on monolingual corpora to build MT systems (Artetxe et al., 2019; Lample et al., 2017; Artetxe et al., 2017). In this scenario, this work aims at generating bilingual resources for those language pairs and domains lacking sufficient parallel corpora by leveraging recent research carried out in the field of unsupervised learning; deriving bilingual dictionaries and translation models from large amounts of monolingual corpora only. Partners include two research centers: the IXA group from the University of the Basque Country and the Barcelona Supercomputing Center; and three companies providing translation services: Tilde, Iconic and Unbabel. The contributions of the project are publicly available[1] and can be summarized as follows:

- Release of 18 monolingual corpora for specific domains and a variety of languages in which both low-resource and high-resource languages are found.
- Release of bilingual dictionaries, word embeddings and translation models.
- Expansion of unsupervised techniques (Artetxe et al., 2019) to further domains and languages, including some low-resourced languages.

The remaining of this article is organized as follows. In Section 2, we summarize the relevant related work. In Section 3, we describe the collected corpora and the experiments. Then, in Section 4, we go through the results of the experiments. Finally, in Section 5, we discuss the obtained results and draw some conclusions.

2. Related Work

The advent of neural machine translation, originally with recurrent sequence-to-sequence models (Sutskever et al., 2014; Bahdanau et al., 2016) and more recently with the renowned Transformer architecture (Vaswani et al., 2017), together with the development of purely unsupervised machine translation algorithms (Artetxe et al., 2019; Lample et al., 2017; Artetxe et al., 2017), offered the promise of revolutionizing the machine translation scene.

Indeed, effective unsupervised machine translation offers a solution to pairs for which little to no parallel corpora are available, as in the case of most of the language pairs and domains targeted in this work (listed in Table 2). Aside from commercial, general-purpose systems (with potentially sub-optimal performance in specialised domains and low-resource languages) such as Google Translate, few systems target very low-resource scenarios and are publicly available. Bawden et al. (2019) in the case of biomedical English-Spanish, or Costa-jussà et al. (2014) in the case of general domain English-Catalan translation are instances of such open-source systems. Other

[1] https://elrc-share.eu/repository/search/?q=MT4all

than that, OpusMT (Tiedemann and Thottingal, 2020), a Marian-based (Junczys-Dowmunt et al., 2018) translator trained on the Opus parallel corpus (Tiedemann and Nygaard, 2004) provides general domain baselines for a number of language pairs, including English-Basque, English-Catalan, English-Kazakh, English-Finnish, English-Norwegian, and English-Latvian. All these works use supervised approaches, thus leaving room for experimenting with the application of unsupervised techniques, such as Artetxe et al. (2019).

Nevertheless, several challenges and questions remain open when applying these unsupervised translation techniques to real use cases involving specialised domains and low-resource languages. In seminal works of unsupervised machine translation, authors used standard benchmarks in machine translation (for the sake of easing comparisons with other methods). They applied the method to non low-resource scenarios, e.g. general domain English-French and English-German, which is an unrealistic unsupervised scenario (Artetxe et al., 2020). This calls for testing the method (Marchisio et al., 2020) in real-world low-resource pairs, for example involving agglutinative languages such as Finnish. In this work, we investigate the application of Artetxe et al. (2019)'s method to an ample range of domains and low-resource pairs, and release the corresponding outcomes as open resources.

3. Method and resources

We train 12 translation models for 10 language pairs and for 6 different domains, the resulting combinations are shown in table 2.

In this section, we describe the collection of corpora to train the models as well as the resulting resources of the training: the translation models, the word embeddings and the dictionaries.

3.1. Training Corpora

We have collected monolingual corpora for 11 languages from a wide range of language families and writing scripts, including Basque, Catalan, English, Finnish, Georgian, German, Kazakh, Latvian, Norwegian, Spanish and Ukrainian; and from 6 different domains, namely Biomedical, Customer support, Finance, General, Legal and Newswire.

The data used for training the machine translation models was obtained from various sources, as shown in Table 1. The collected corpora comes mainly from (1) data crawled within the scope of this work, which will be openly released, (2) OSCAR (Ortiz Suárez et al., 2019), an open-source collection of monolingual corpora extracted from the web for different languages[2], and (3) other domain and language-specific publicly available corpora.

For **Basque**, we have also employed the Elhuyar Web Corpus (Leturia, 2012), a corpus of Basque collected from the web, and IXA's own crawling, extracted from a controlled crawling of newspaper websites.

The corpus of **Catalan** corresponds to the *Catalan Textual Corpus* (Armengol-Estapé et al., 2021). It includes already publicly available datasets (among which OSCAR is included) and recently crawled data.

The sources of the crawled data for the **financial** domain include bank websites, finance resource sites, finance blogs and forums on banking and economy-related issues.

The sources of the crawled data for the **legal** domain include legislation websites, governmental sites, and domains from the Court and the Parliament.

The sources of the crawled data for the **customer support** include FAQ and help websites, as well as community sites and forums. Furthermore, a general domain corpus crawling for English, German, Norwegian and Spanish has been created by crawling the URLs under the superdomain of the targeted subdomains.

The sources of the **biomedical** domain corpus vary for each language. In the case of Spanish, we used CoWese (Carrino et al., 2021), a crawling of urls from the biomedical domain. The English corpus for this domain consists of the English counterpart of the MeSpEn corpus (Villegas et al., 2018), as well as the UFAL corpus[3]. Also, we included the English part of the EMEA (Tiedemann, 2012) and BARR2 (Intxaurrondo et al., 2018) datasets.

Finally, for **English**, besides our crawlings and OSCAR, we also used the News-crawl corpus[4], which gathers news articles in English. For training, we used news from 2007 until 2013.

3.1.1. Corpora Preprocessing

To improve the quality of the crawled data, we use CorpusCleaner (Armengol-Estapé et al., 2021), a preprocessing pipeline tool designed to obtain clean raw text from crawled data. This pipeline supports all the languages targeted. The pipeline is based on a data parser, several language identifiers, different filters based on heuristics as well as deduplication mechanisms both at sentence and document level.

We have used CorpusCleaner also on the corpora obtained from the OSCAR website. For English, German and Spanish, we only use a random subset of 45 million documents (in OSCAR, each sentence is considered a document) as the whole OSCAR would be too large for our purposes. The resulting statistics of the collected and processed corpora are shown in table 1.

3.2. Translation Models

We have trained MT models between English and 12 combinations of languages and domains: generic Basque and Catalan; financial Norwegian, Finnish, Latvian; biomedical Spanish; legal Ukrainian, Kazakh, Georgian; and finally Norwegian, Spanish and German, in the domain of customer support.

[2]We use OSCAR in all systems training except in the biomedical domain system training.

[3]https://ufal.mff.cuni.cz/ufal_medical_corpus

[4]https://data.statmt.org/news-crawl/en/

	Domain	Source	Sentences	Tokens
Basque (EU)	General	Elhuyar Web Corpus	7,881,727	120,668,124
	General	MT4All	7,448,381	106,857,055
	General	OSCAR	2,955,770	44,327,996
Catalan (CA)	General	Catalan Textual Corpus	73,172,152	1,758,388,896
English (EN)	Biomedical	MT4All	39,127,771	601,611,211
	Customer Support	MT4All	169,344	2,025,813
	Financial	MT4All	3,672,407	67,732,742
	Legal	MT4All	141,389	2,465,121
	General	MT4All	5,917,753	93,445,485
	General	OSCAR	108,546,661	2,284,911,064
	Newswire	News-crawl	44,999,975	908,660,388
Finnish (FI)	Financial	MT4All	3,624,828	44,613,629
	General	OSCAR	112,927,936	1,491,094,267
Georgian (KA)	Legal	MT4All	189,506	3,688,666
	General	OSCAR	5,032,382	88,047,553
German (DE)	Customer Support	MT4All	87,002	1,315,050
	General	MT4All	4,880,000	67,637,441
	General	OSCAR	157,751,465	1,989,472,063
Kazakh (KK)	Legal	MT4All	119,711	1,862,857
	General	OSCAR	9,129,117	98,868,498
Latvian (LV)	Financial	MT4All	485,845	8,827,703
	General	OSCAR	13,778,938	225,211,959
Norwegian (NO)	Financial	MT4All	5,224,308	93,523,653
	Customer Support	MT4All	30,757	490,559
	General	MT4All	2,692,915	43,424,915
	General	OSCAR	31,090,926	626,718,236
Spanish (ES)	Biomedical	CoWeSe	41,236,605	919,783,046
	Customer Support	MT4All	58,490	1,054,268
	General	MT4All	895,644	16,725,511
	General	OSCAR	131,866,954	2,225,124,380
Ukrainian (UK)	Legal	MT4All	7,544,396	69,128,091
	General	OSCAR	84,502,198	1,000,763,332

Table 1: Statistics of the collected corpora per language and domain

We have relied on Monoses (Artetxe et al., 2019) to train translation systems based exclusively on monolingual corpora. This section briefly summarises the training process behind this tool. The process mainly consists of four steps: (1) generation of monolingual word embeddings, (2) generation of bilingual word embeddings via linear mapping, (3) inference of a Statistical Machine Translation (SMT) system and training of a Neural Machine Translation (NMT).

The process begins by generating independent word embeddings for each language based on monolingual corpora. Embeddings are continuous representations of words based on distributional semantics which allow, among other things, measuring the semantic similarity between words. However, the embeddings generated separately for each language are independent and cannot be directly compared. So, the next step is to align these embeddings to generate bilingual embeddings, where the words of both languages are represented in a joint vector space. This process is carried out through an iterative procedure that seeks to minimize the distance of each word in one language with the closest word in the other language, thus finding the linear transformation that best combines both spaces. To avoid getting stuck in a bad quality local minimum, the process needs an initialisation, which can be created automatically based on cognates found in both vocabularies and exploiting the structural similarity of the embeddings.

Furthermore, these embeddings can also be used to generate translation models without the need of parallel corpora. In this task, the modular architecture of Statistical Machine Translation (SMT) facilitates the training of the system, since, with the exception of the translation model, the rest of the required models can be trained without the need for bilingual corpus. The translation table, in contrast, can be generated in an analogous way to bilingual dictionaries, that is, keeping the best n translations for each word and assigning to them a translation probability based on the semantic similarity of each translation. In this way we would already have a translation system that is the starting point of an iterative process in which the SMT system is retrained, using this time a synthetic corpus generated automatically through back-translation (Sennrich et al., 2015). After a couple of iterations in which the SMT system is improved, a dual Neural Machine Translation (NMT) system is trained. Dual NMT uses the au-

tomatically generated translations in one direction as synthetic corpora to train the system in the opposite direction.

In the case of languages with rich morphology that have obtained poor results with the usual training (Basque and Finnish), we have modified the training process to apply Byte-Pair Encoding (BPE) (Sennrich et al., 2016) from the beginning of the process instead of only in the training of final NMT systems. That is, the corpus is segmented before generating the monolingual embeddings, which in this case would be subword embeddings. This reduces the number of unknown words in the generation process of the SMT models, which has been found to be a problem for these languages. In this work, we have focused on the application of the BPE technique to those morphologically rich languages that have obtained poor results, leaving its application to other languages and domains for future work.

When training the models, we have used two hardware configurations. On the one hand, we used an IBM Power9 8335-GTH with 160 threads for SMT training and 4 GPUs NVIDIA V100 (Volta) with 16GB HBM2 for NMT training. On the other hand, we employed an Intel Xeon E5-2687W v3 with 40 threads for SMT training and 4 GPUs NVIDIA Titan V with 12 GB for NMT training. The average time taken to train the model when using the first configuration was 10 days, while using the second configuration the time taken for training was 17 days.

3.3. Embeddings and Dictionaries

In this work, we have created bilingual word embeddings and dictionaries for each of the language pairs selected to train the models mentioned in subsection 3.2.

On the one hand, in order to create bilingual word embeddings we have made use of the Vecmap tool (Artetxe et al., 2018). The goal of the VecMap is to learn the linear transformation matrices so the mapped embeddings are in the same cross-lingual space. The method consists of four sequential steps. Firstly, VecMap length normalizes the embeddings, then mean centers each dimension, and then length normalizes them again. Secondly, the initialization builds a dictionary based on similarity matrices. Thirdly, with the initial dictionary, the self-learning iterates the following two steps: (1) compute the optimal orthogonal mapping maximizing the similarities for the current dictionary and (2) induce a new dictionary based on the similarity matrix of the mapped embeddings. Finally, the refinement step further improves the resulting mapping through symmetric re-weighting (Artetxe et al., 2018).

On the other hand, to create the dictionaries we have employed TermSuite [5]. TermSuite is a toolbox for terminology extraction and multilingual term alignment. Firstly, we have extracted the terminology of the English corpora we have used to train the models. Thereafter, we find those English terms in the translation table that has been created during the SMT model training. Finally, with the aim of choosing the best options, we have filtered the possible candidates and selected the ones with a frequency higher than 10 and a probability higher than 0.25.

4. Evaluation

In this section, we describe the test sets that have been used in order to evaluate the models, as well as the metrics used for automatic evaluation. Lastly, we report the automatic evaluation results and the human evaluation results.

4.1. Test sets

The test sets used for evaluation originate from (1) already publicly available MT test sets or (2) brand new creations purposely built for the project, as shown in Table 2.

From the already published test sets, for **Catalan**, we use the Catalan United Nations test set (Costa-jussà, 2020), which is the Catalan translation of the United Nations Parallel Corpus test set (Ziemski et al., 2016). In the case of the **biomedical** domain, we make use of the WMT 20 Biomedical Shared Task test set (Bawden et al., 2020). We have produced new test sets for all other language pairs.

For **Basque**, we use the Basque translation (Altuna et al., 2017) and its original English dataset of the MEANTIME corpus (Minard et al., 2016). This English dataset is composed by a total of 119 English WikiNews[6]. It consists of news articles on economics topics such as Airbus and Boeing, Apple Inc., Stock market, and General Motors, Chrysler and Ford. The Basque translation was translated from scratch and manually aligned to the original English version.

For the **customer support** domain, test sets were compiled by browsing the web for freely available English language customer service email templates. In total, 165 emails were compiled, with an average length of 95 words. These English emails were then reviewed internally and URLs, people, product and company names were edited or added to ensure there was enough named entity variety. The emails were then sent to professional linguists to translate from scratch into Norwegian, German and Spanish.

For the **financial** domain, the test sets were created by crawling parallel data from financial institution websites that publish content in the two languages. The crawled data was then manually checked by professional translators. Non-parallel segments were discarded and only parallel segments remained.

The test sets for the **legal** domain were collected using two different methods. Some were gathered by browsing public websites from institutions and governments in the relevant countries and their official translations

[5] https://github.com/termsuite/termsuite-core

[6] http://en.wikinews.org/

into English, and others were internal tests sets collected over the years which contain a mix of public and proprietary data.

4.2. Metrics

A basic evaluation protocol was established to ensure that all partners would use the same metrics to evaluate the engines trained within the project. Thus, even though some partners may have used additional evaluation metrics and/or human quality analysis, all engines were at least evaluated with the same state-of-the-art automatic metric, namely sBLEU (Post, 2018).

4.3. Automatic Evaluation Results

The results of the automatic evaluation are shown in table 3.

As can be observed, the unsupervised models do not achieve parity with the supervised model, which is to be expected, although, remarkably, in some cases the results are not too far away. It is important to note that unsupervised systems trained without any kind of bilingual information cannot be expected to significantly outperform supervised systems that have access to previously translated sentences.

If we focus on the results obtained, in some cases the unsupervised and supervised systems are close, such as in the case of the English→Spanish MT system developed for the customer support domain (30.3 vs. 34.6 sBLEU points respectively). We also observe that the use of BPE from the start of the training process, although improving the results in the two morphologically rich languages, is more effective in boosting the score for English↔Finnish (11.1 → 17.5 and 8.8 → 21.6) than for English↔Basque (5.1 → 9.0 and 12.1 → 16.0).

If we analyze the results of the unsupervised models in more depth, it becomes clear that the difference between the results of the different language pairs is modelled by several factors: the size of domain-specific corpora available for each chosen domain, the number of general corpora available, the distance of the languages from English, and their level of complexity. For example, the results for the English↔Ukrainian language pair for the legal domain can be considered normal for an unsupervised system in this particular language direction. The lack of data in the legal domain to train the system is somewhat compensated for by a significant amount of general domain data (more than 80 million sentences for both languages).

Looking at the results of English→Georgian and English→Kazakh the evaluation results of the unsupervised systems are much lower than those obtained by supervised systems, and can be considered bad even in an unsupervised setting. In these cases both the general domain corpora (less than 10 million phrases for both languages) and legal domain corpora (less than 3 million phrases for both languages) for Georgian and Kazakh are clearly insufficient to obtain acceptable results.

As for the results obtained by the unsupervised systems trained in the customer support domain, the results are somewhat more positive. For the English→German language pair the unsupervised model obtains 30.6 sBLEU points, while Google Translate scores 56 sBLEU points. Although the difference is significant, we must bear in mind that there are many bilingual corpora available for this particular language pair and that it is one of the most widely used language pairs in the world in terms of translation. It is interesting to note that the English→Norwegian system obtains somewhat similar results (28.9 sBLEU points) to those obtained by the English→German system (30.6 sBLEU points) even though it was trained using a far smaller general corpus (33 million sentences versus 157 million sentences) and smaller domain-specific corpus (30K sentences versus 87K sentences). In contrast to the English→German language pair, these results are closer to those obtained by the supervised systems due to the relative greater difficulty of obtaining large parallel corpora for this language pair.

Finally, the best results in terms of the benchmarking against supervised systems have been obtained by the English→Spanish unsupervised models, both for the customer support domain and for the biomedical domain. However, the way in which these two systems have been trained is quite distinct. While the system trained for the customer support domain has been trained with fewer domain-specific sentences (less than 1 million) and a large general domain corpus (over 131 million) in Spanish, the system developed for the biomedical domain has been trained using only biomedical domain corpora (over 39 million sentences for English and over 41 million sentences for Spanish). The highest scores have been achieved by the system trained for the biomedical domain (41.6 sBLEU points for English→Spanish, and 39.7 sBLEU points for Spanish→English). This could likely be attributed to the fact that the MT system was able to adequately adapt to the biomedical domain on account of the large amount of domain-specific data that it was trained on.

In the interest of further comparison, we have compared the unsupervised system with the WMT 2020 Biomedical Translation Shared Task (Bawden et al., 2020) participating systems that have been trained with data from the biomedical domain. As Google Translate is trained on general data, we thought that this might give us a clearer indication of the performance of the unsupervised model. Upon comparison of the unsupervised system against supervised systems trained for the biomedical domain, the difference between the systems is small. The best result for the English→Spanish direction was obtained by the Sheffield system (Soares and Vaz, 2020) in its unique run (44.93 sBLEU points), while the results of the other systems ranged from 37.55 to 44.64 sBLEU points. In the opposite direction (Spanish→English), the best system was NLE (Naver Labs Europe) in its run 1 (50.57 sBLEU points) and

Translation pair	Domain	Source	Sentences	English Tokens	L2 Tokens
EN↔ES	Biomedical	WMT20 Shared task	1,009	20,095	23,007
EN↔DE	Customer Support	MT4ALL	1,668	15,634	16,104
EN↔ES	Customer Support	MT4ALL	1,668	15,634	14,272
EN↔NO	Customer Support	MT4ALL	1,688	15,634	16,298
EN↔LV	Financial	MT4ALL	1,000	17,035	13,428
EN↔FI	Financial	MT4ALL	1,000	11,257	7,517
EN↔NO	Financial	MT4ALL	1,000	17,318	14,942
EN↔CA	General	UN Test Set	4,000	106,733	123,696
EN↔EU	General	MT4ALL	1,788	34,747	28,597
EN↔KA	Legal	MT4ALL	1,099	25,172	18,795
EN↔KK	Legal	MT4ALL	1,000	15,822	12,495
EN↔UK	Legal	MT4ALL	964	19,318	16,73

Table 2: Evaluation sets for each language pair and domain

		sBLEU		
Languages	Domain	MT4ALL	BPE	GT
EN → FI	Financial	11.1	17.5	28.3
FI → EN	Financial	8.8	21.6	36.1
EN → LT	Financial	24.7	-	35.7
LT → EN	Financial	15.1	-	36.9
EN → NO	Financial	27.5	-	36.3
NO → EN	Financial	23.4	-	40.4
EN → CA	General	30.8	-	39.7
CA → EN	General	33.4	-	47.4
EN → EU	Newswire	5.1	9.0	19.7
EU → EN	Newswire	12.1	16.0	31.8
EN → UK	Legal	14.2	-	28.2
UK → EN	Legal	15.4	-	28.7
EN → KA	Legal	12.0	-	24.9
KA → EN	Legal	18.6	-	39.6
EN → KK	Legal	6.4	-	20.4
KK → EN	Legal	7.7	-	26.7
EN → DE	Customer Support	30.6	-	56.0
DE → EN	Customer Support	35.2	-	43.8
EN → ES	Customer Support	30.3	-	34.6
ES → EN	Customer Support	33.3	-	42.1
EN → NO	Customer Support	28.9	-	41.3
NO → EN	Customer Support	31.4	-	37.9
EN → ES	Biomedical	41.6	-	48.6
ES → EN	Biomedical	39.7	-	49.1

Table 3: Automatic metric results for all tested language directions and domains reported in sBLEU. MT4ALL refers to our model, BPE refers to our model with BPE segmentation and GT refers to Google Translate

the other systems ranged from 19.93 to 46.34 sBLEU points.

4.4. Human Evaluation Results

In some cases, the partners in the project decided to carry out a human evaluation of the output of the unsupervised MT systems trained in the project. This was the case of the customer support, financial and biomedical domain.

4.4.1. Customer Support domain

In the customer support domain, a cursory inspection of the outputs shows fluency and accuracy problems that are particularly critical in the customer service domain. Across all three language pairs, but especially for EN→NO and EN→ES, we see the unsupervised systems mistranslating personal names in customer service emails (e.g.: '*Barney*' → '*Rosendo*' (EN→ES); '*John*' → '*Olav*' (EN→NO); '*Mike*' → '*Holger*' (EN→DE), a serious error that is likely to produce a negative perception in the final user of the email.

Unlike their commercial counterparts, the unsupervised systems did not localize dates with the YYYY-MM-DD or the MM-DD-YYYY format (in all three language pairs the preferred format is DD[-/]MM[-/]YYYY). Properly localizing the dates in customer service emails is critical, especially when they make reference to warranty expiration dates, as not doing so can lead to confusion in the end user.

Other problems present in the output of the unsu-

pervised system involve alternating between the use of formal and informal forms in the same texts in EN→ES and EN→DE; duplications (e.g.: Thanks → Takk, takk (EN→NO)) and parsing of URLs (e.g.: ('*www.acme.no/blog*' → '*www.acme.no / blogg*'). The latter can seriously hinder the usefulness of the email.
Neither the unsupervised systems nor the commercial supervised systems, working on a sentence level, could handle document-level phenomena that are important in customer service, again adding to the negative perception it may have in the final user, which in turn has a cultural component. In EN→DE, for example, the first sentence after the greeting must be lower-cased, but none of the systems got this right.

4.4.2. Financial domain

For the financial domain, we carried out a small-scale error analysis of the EN→LV MT system to better understand the quality of the unsupervised NMT system and the types of errors the system makes. We chose this language pair since we had direct access to human annotators proficient in both languages.
The error analysis was performed using the Translation Quality Assessment feature of the SDL Trados Studio[7] computer-assisted translation tool with the MQM Core project template. We analysed a random subset of 100 sentences from the test set.
In total, we found 135 errors in the translations. The results of the error analysis are depicted in Table 4 and show that the unsupervised MT system:

- Introduced quite a few mistranslation issues (44 out of 135 errors), of which the majority (28) are marked as major errors. Most of these errors were related to bad lexical choices. In three cases, the subject and object were swapped.
- Sometimes omitted phrases. This was especially evident for named entities, where nine out of 17 omission errors were related to named entities being partially or fully omitted.
- Introduced quite a few terminology errors (18 in total), although it is adapted to the financial domain. These can also be considered as errors relating to poor lexical choices.
- Introduced quite a few grammar-related errors (19 in total), of which 16 were related to wrong inflections, including inaccurate agreements between constituents due to incorrect inflections. Two errors were related to missing or erroneous prepositions, and one error was related to incorrect word order.
- Introduced 13 errors related to typography, of which seven errors were related to incorrect capitalisation of words, three errors were related to usage of the incorrect dash symbol, two errors were related to extra/unnecessary spaces, and the remaining error was related to an unnecessary comma.
- Introduced 13 errors that hinder the understanding of the translation (all minor errors, however). These errors were related mostly to translations that are too literal, and word order that is incorrect and does not read fluently.
- Presented another 11 errors, out of which five errors were related to untranslated phrases, four errors were related to locale convention errors, one error was related to inconsistent translation, and the last error was related to stylistics.

[7] https://www.trados.com/products/trados-studio/

Error category	Minor (1)	Major (5)	Critical (10)	Error score
Accuracy				
Mistranslation	14	28	2	174
Omission	7	9	1	62
Terminology	16	2	0	26
Untranslated	2	3	0	17
Fluency				
Inconsistency	1	0	0	1
Stylistics	1	0	0	1
Grammar	18	1	0	23
Locale convention	3	1	0	8
Typography	10	3	0	25
Unintelligible	13	0	0	13
Total	85	47	3	350

Table 4: Results of the MQM-based error analysis of the EN→LV unsupervised MT system (error categories with no errors were omitted)

According to the MQM methodology adopted in the SDL Trados Studio CAT tool, the 100 sentences translated by the unsupervised MT system received a penalty score of 0.22 and a total error score of 350 per 1000 words. The default Pass threshold for professional translators in the MQM configuration of SDL Trados Studio is 50 per 1000 words. This means that the unsupervised MT methods would require further research and improvements to reduce the gap between translation quality.
The unsupervised MT system introduced errors in 67 sentences (1.35 errors per sentence). On a more positive note, the unsupervised MT system did manage to translate 33 out of the 100 sentences without any errors, which shows that there is potential for this technology in scenarios where there is no, or very little, parallel data available.
To better understand the cause of the relatively large number of accuracy errors, we analysed the quality of the bilingual dictionaries that were generated as intermediate resources for the unsupervised MT system. For this, we selected a random sub-set of 100 entries in the dictionary and evaluated the parallelism of the entries. The analysis showed that 69% of entries were correct, 11% consisted of unrelated source

and target words and phrases (e.g., '*Cuban*' and '*popular*'), 10% were pairs representing different terms from the same domain (e.g., '*dealership*' and '*service*'), 4% consisted of similar but still different words (e.g., '*presentable*' and '*attractive*'), 4% were partial translations or overlapping (invalid alignment) translations (e.g., '*time limit*' and '*limit*'), and 2% were entries that contained words and phrases with opposite meanings (e.g., '*domestically*' and '*abroad*'). The high error rate of the dictionary (i.e., 31%) is a likely cause for the accuracy-related translation errors.

4.4.3. Biomedical domain

Finally, in the case of the biomedical EN→ES model, an MQM analysis was carried out, revealing similar types of errors as the ones observed in the human inspection carried out for the customer domain. The analysis was performed on the complete test set for the language direction EN→ES by a professional native translator.

We followed the same methodology than for the financial domain. Regarding accuracy, the unsupervised system introduced many critical entity mistranslations and terminology issues, to be expected from unsupervised models, and in a specific domain. Also, a high number of addition errors were found, since the model tends to repeat words. Regarding fluency, most errors are minor spelling issues and there are almost no critical errors for fluency, which reinforces the idea that neural translations sound more natural than other methods.

The 1009 sentences translated by the unsupervised MT system received a penalty score of 0.14 and a total error score of 138.09 per 1000 words, which is much lower than in the case of the financial domain. If we take as reference the default Pass threshold for professional translators which is 50, this would actually be considered a positive result as we are dealing with an unsupervised system. The unsupervised MT system introduced 603 errors in 1,009 sentences (0.59 errors per sentence) and managed to translate 519 out of the 1,009 sentences without any errors, which shows that there is high potential for this technology, especially when having large amounts of data.

5. Discussion and Conclusion

One of the most remarkable aspect of the achieved results is the variability of the scores obtained for each translation pair. Even if the results for pairs such as Ukrainian↔English and Georgian↔English are below 20 sBLEU points and below 10 sBLEU points for Kazakh↔English, the scores reached by the three languages pairs in the customer support domain (German↔English, Norwegian↔English and Spanish↔English) reach around 30 sBLEU points, and around 40 sBLEU points in the biomedical domain.

It is worth noting the positive impact of the size of the in-domain corpus as seen in the biomedical domain for the Spanish↔English results. Thus, the total size of the corpus used to train the biomedical models is considerably smaller than the full corpus used in the customer support model (around 41M sentences vs. around 133M sentences for Spanish and around 39M sentences vs. around 153M sentences for English), but the former obtains 11 to 6 sBLEU additional points. The reason for this is that the biomedical corpus is fully in-domain while the customer support model was built using 132M sentences of general domain data and only around 1M sentences of in-domain data.

While the results shown by automatic metrics for the Spanish biomedical domain system were already positive, the human evaluation not only confirmed this, but also showed that the engine was good for production environments. In fact, the human evaluation revealed that the quality of the unsupervised system was higher than what the automated metrics seemed to suggest.

We have also observed an interesting fact related to the two language directions involving Basque, as they achieve extremely different scores (5.1 vs. 12.1 sBLEU points without BPE). Although the linguistic features of the languages involved may partly justify some differences, this difference is unexpectedly high, as the system that translates in one direction is used to train the opposite direction, and typically the quality of both systems tends to progress in parallel. We also observe a similar behaviour in other language pairs, both low-scoring, like Georgian↔English (12 vs. 18.6 sBLEU points), and with good results, like German↔English (30.6 vs. 35.2 sBLEU points). The three languages involved have a rich morphology, which may influence derive in lower results when translating from English.

Finally, if we compare the results achieved with Ukrainian and Georgian in the legal domain we see that they both reach similar results as measured in sBLEU, but the size of the training corpus is vastly different (around 5M sentences vs. 92M sentences).

The conclusion that we can derive from our results is that in low-resource scenarios, completely unsupervised systems tend to yield poor results, except when the amount of in-domain monolingual data is big enough to compensate. Moreover, while in our evaluation exercise, we have purposely adopted a purely unsupervised approach. Nonetheless, the fact is that in real-world scenarios there is always some access to parallel data or it can be created synthetically by triangulation or other methods. Therefore, a more reasonable use of unsupervised techniques is to combine them in a semi-supervised manner with some amounts of parallel data. Finally, the monolingual corpora presented in this paper and the generated bilingual resources can be considered a positive contribution of this work.

6. Acknowledgements

This work was funded by the MT4All CEF project.[8]

[8] https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom/2019-eu-ia-0031

7. Bibliographical References

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Artetxe, M., Labaka, G., and Agirre, E. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July. Association for Computational Linguistics.

Artetxe, M., Ruder, S., Yogatama, D., Labaka, G., and Agirre, E. (2020). A call for more rigor in unsupervised cross-lingual learning. *CoRR*, abs/2004.14958.

Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bawden, R., Bretonnel Cohen, K., Grozea, C., Jimeno Yepes, A., Kittner, M., Krallinger, M., Mah, N., Neveol, A., Neves, M., Soares, F., Siu, A., Verspoor, K., and Vicente Navarro, M. (2019). Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy, August. Association for Computational Linguistics.

Bawden, R., Di Nunzio, G., Grozea, C., Unanue, I., Yepes, A., Mah, N., Martinez, D., Névéol, A., Neves, M., Oronoz, M., et al. (2020). Findings of the wmt 2020 biomedical translation shared task: Basque, italian and russian as new additional languages. In *5th Conference on Machine Translation*.

Costa-jussà, M. R., Fonollosa, J. A. R., Mariño, J. B., Poch, M., and Farrús, M. (2014). A large spanish-catalan parallel corpus release for machine translation. *Comput. Informatics*, 33:907–920.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Leturia, I. (2012). Evaluating different methods for automatically collecting large general corpora for basque from the web. In *Proceedings of COLING 2012*, pages 1553–1570.

Marchisio, K., Duh, K., and Koehn, P. (2020). When does unsupervised machine translation work? *CoRR*, abs/2004.05516.

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, Brussels, Belgium, October. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sennrich, R., Haddow, B., and Birch, A. (2016). 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, August. Association for Computational Linguistics.

Soares, F. and Vaz, D. (2020). Uos participation in the wmt20 translation of biomedical abstracts. In *Proceedings of the Fifth Conference on Machine Translation*, pages 870–874.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

8. Language Resource References

Altuna, B., Aranzabe, M. J., and de Ilarraza, A. D. (2017). Eusheideltime: Time expression extraction and normalisation for basque. *Procesamiento del Lenguaje Natural*, (59):15–22.

Armengol-Estapé, J., Carrino, C. P., Rodriguez-Penagos, C., de Gibert Bonet, O., Armentano-Oller, C., Gonzalez-Agirre, A., Melero, M., and Villegas, M. (2021). Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online, August. Association for Computational Linguistics.

Carrino, C. P., Armengol-Estapé, J., Bonet, O. D. G., Gutiérrez-Fandiño, A., Gonzalez-Agirre, A., Krallinger, M., and Villegas, M. (2021). Spanish biomedical crawled corpus: A large, diverse dataset

for spanish biomedical language models. *CoRR*, abs/2109.07765.

Costa-jussà, M. R. (2020). Catalan united nations v1.0 test set, June. This work is supported by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund, through the postdoctoral senior grant Ramón y Cajal.

Intxaurrondo, A., Marimón, M., González-Agirre, A., Lopez-Martin, J. A., Rodriguez, H., Santamaria, J., Villegas, M., and Krallinger, M. (2018). Finding mentions of abbreviations and their definitions in spanish clinical cases: The barr2 shared task evaluation results. *IberEval@ SEPLN*, 2150:280–289.

Minard, A.-L., Speranza, M., Urizar, R., Altuna, B., Van Erp, M., Schoen, A., and Van Son, C. (2016). Meantime, the newsreader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Tiedemann, J. and Nygaard, L. (2004). The opus corpus-parallel and free: http://logos. uio. no/opus. In *LREC*. Citeseer.

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

Villegas, M., Intxaurrondo, A., Gonzalez-Agirre, A., Marimon, M., and Krallinger, M. (2018). The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. *LREC MultilingualBIO: multilingual biomedical text processing*.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia, May. European Language Resources Association (ELRA).