Neural Machine Translation of clinical texts between long distance languages

Xabier Soto
Faculty of Informatics, Ixa Research Group,
University of the Basque Country (UPV/EHU).
Manuel Lardizabal 1, 20018, Donostia (Spain).
xabier.soto@ehu.eus, +34943015110.

Olatz Perez-de-Viñaspre, Gorka Labaka, Maite Oronoz
Computer Languages and Systems, Ixa Research Group,
University of the Basque Country (UPV/EHU), Donostia (Spain)

Keywords
Neural Networks – Machine Translation – Electronic Health Records

Words: 4230

This is a pre-copyedited, author-produced version of an article accepted for publication in **Journal of the American Medical Informatics Association**, following peer review. The version of record Xabier Soto, Olatz Perez-de-Viñaspre, Gorka Labaka, Maite Oronoz, *Neural machine translation of clinical texts between long distance languages*, **Journal of the American Medical Informatics Association**, Volume 26, Issue 12, December 2019, Pages 1478–1487, is available online at: https://doi.org/10.1093/jamia/ocz110

Abstract

Objective

To analyze techniques for Machine Translation of Electronic Health Records (EHRs) between long distance languages, using Basque and Spanish as a reference. We studied distinct configurations of Neural Machine Translation (NMT) systems and used different methods to overcome the lack of a bilingual corpus of clinical texts or health records in Basque and Spanish.

Materials and Methods

We trained Recurrent Neural Networks (RNN) on an out-of-domain corpus with different hyperparameter values. Subsequently, we used the optimal configuration to evaluate machine translation of EHR templates ~~in~~ between Basque and Spanish, using manual translations of the Basque templates into Spanish as a standard. We successively added to the training corpus clinical resources, including a Spanish-Basque dictionary derived from resources built for the machine translation of the Spanish edition of SNOMED CT into Basque, artificial sentences in Spanish and Basque derived from frequently occurring relationships in SNOMED CT, and Spanish monolingual EHRs. Apart from calculating BLEU values, we tested the performance in the clinical domain by human evaluation.

Results

We achieved slight improvements from our reference system by tuning some hyperparameters using an out-of-domain bilingual corpus, obtaining 10.67 BLEU points for Basque-to-Spanish clinical domain translation. The inclusion of clinical terminology in Spanish and Basque and the application of the backtranslation technique on monolingual EHRs significantly improved the performance, obtaining 21.59 BLEU points. This was confirmed by the human evaluation performed by two clinicians, ranking our machine translations close to the human translations.

Discussion

We showed that, even after optimizing the hyperparameters out-of-domain, the inclusion of available resources from the clinical domain and applied methods were beneficial for the described objective, managing to obtain adequate translations of EHR templates.

Conclusion

We have developed a system which is able to properly translate health record templates from Basque to Spanish without making use of any bilingual corpus of clinical texts or health records.

INTRODUCTION AND MAIN OBJECTIVE

Our objective is to analyze different techniques for Basque-to-Spanish and Spanish-to-Basque Machine Translation (MT) in the clinical domain. Specifically, distinct

configurations of Neural Machine Translation (NMT) systems were tested leveraging the limited resources available for the clinical domain in Basque and Spanish.

Basque is a minoritised language, sharing a bilingual environment with the strong language Spanish. This is reflected in the Basque public health service, where nearly all of the health records are registered in Spanish so that any doctor can understand them. Nowadays, if any patient wants to consult their health record in Basque, it is translated on demand by human translators, the translation is given to the patient, and the public health service does not retain a copy. With a view to guaranteeing the linguistic rights of all doctors and patients, our purpose is to develop a NMT system so that Basque speaking doctors are able to write in Basque and patients can read their health records in the language of their choice without waiting for a manual translation.

The increasing availability of Electronic Health Records (EHR) makes the application of advanced MT techniques possible. However, our main handicap is the lack of bilingual corpora for the clinical domain in Basque and Spanish. To alleviate this problem, different approaches were tried such as i) inserting a medical bilingual dictionary to an out-of-domain corpus, ii) creating artificial sentences from the relations in SNOMED CT, iii) adding a clinical domain monolingual corpus, along with its backtranslation or, iv) using the repetition of the monolingual corpus as if it was bilingual.

As a sample of the results that will be presented further, Table 1 shows an example of a sentence translated by a bilingual doctor, by Google Translate [1], our system using only out-of-domain corpora, and our system including SNOMED CT terminology.

Original sentence in Basque
lipido-en *metabolismo-aren* *asaldura*
lipid-GEN.PL metabolism-GEN.SG disorder
'disorder of lipid metabolism'
Manual translation into Spanish
trastorno *metabolismo* *lípido-s*
disorder metabolism lipid-PL
'disorder metabolism lipids'
Translation by Google Translate
metabolismo *de* *los* *trastorno-s* *lipídico-s*
metabolism of the.M.PL disorder-PL lipid-PL
'metabolism of the lipid disorders'
Translation by the system trained with the out-of-domain corpus
Alteración *de-l* *metabolismo* *de* *las* **lipides*
disorder of-the.M.SG metabolism of the.F.PL *lipides
'disorder of lipides metabolism'
Translation by the system trained including SNOMED CT terminology
el *trastorno* *de-l* *metabolismo* *de* *los* *lípido-s*
the.M.SG disorder of-the.M.SG metabolism of the.M.PL lipid-PL
'the disorder of the lipid metabolism'

Table 1: Basque sentence translated into Spanish by a human and by different systems.

Our main contributions are:

- The hyperparameter optimization of a NMT system dealing with long distance languages, including a morphologically rich language.
- A high quality translation of clinical texts without an in-domain bilingual corpus of texts or records, making use of bilingual terminological resources and techniques that leverage specialised lexica and monolingual corpora.

BACKGROUND

Machine Translation is defined as the process of automatically translating a text from one natural language to another. In this work, we focus on Neural Machine Translation (NMT), which is the result of applying the theory of Neural Networks to Machine Translation. The idea was first suggested as early as 1997 [2, 3], but computational limitations did not allow it to be pursued at that time. Fifteen years later, the idea was recovered [4, 5] with real possibilities of applying it.

Neural Networks for Machine Translation usually rely on encoder-decoder configurations, with one neural network for each encoder and decoder. The process of training a Neural Network consists of making a prediction starting with some initial weights, calculating the error according to the training data, and updating the weights of the system using techniques such as backpropagation [6] until some loss function is minimized. Next, the trained model is tested with new data.

The results of NMT systems can vary depending on the architecture, number of layers, number of neurons per layer and other configuration parameters. In order to distinguish them from the parameters (weights and bias) learned during the training process, the base configuration parameters are usually referred to as hyperparameters.

The main characteristic of NMT systems compared to previous techniques is that they act as black boxes that learn to translate without making use of any explicit linguistic or statistical information. To do this, the text in the source language is encoded into numerical values, representing word and sentence meanings as vectors, which are then decoded into sentences in the target language.

Recently, the neural approach has proven to be the most effective for Machine Translation when a large bilingual corpus is available [7], making some significant improvements with the inclusion of an attention-mechanism to automatically search for the most relevant words on a source sentence to be translated into the next output word [8], or using word segmentation to improve the translation of rare words [9].

When approaching our objective, that is, building a NMT system between Basque and Spanish for the clinical domain, there are several perspectives that have to be considered, which can be divided into three areas: NMT between long distance languages; domain adaptation for NMT; and the handicap of performing the NMT task with no in-domain bilingual corpus. In this Section, we mention the relevant works in each of these areas, although in some cases the developed techniques respond simultaneously to more than one of the described problems.

NMT between long distance languages

Spanish is a Latin-derived language sharing characteristics with other European languages, while Basque is a pre-Indo-European language, completely isolated.

Briefly, Basque is a highly agglutinative language, with a rich morphology, where words are usually created adding suffixes that mark diverse cases. The verb morphology is especially complex, including morphemes that add information about the subject, object, number, tense, aspect, etc. Furthermore, the order inside the sentences is relatively free, which makes the development of NMT systems for Basque a challenging task, particularly for evaluation purposes.

Recent work shows that better results can be obtained with NMT for Basque than with the traditional rule-based or statistical techniques [10]. Specifically, Etchegoyhen et al. approach the complex morphology problem by testing different word segmentation methods, from linguistically motivated ones, to the well known Byte Pair Encoding (BPE) word segmentation method. For the sentence order variability, they created a second reference for the test set; and they also tested different values of length-normalization and coverage-penalty, based on previous work [11]. Length-normalization is a hyperparameter that compensates for the tendency of the system to choose sentences of shorter length, as the probability of a given output sentence is calculated by multiplying the probabilites of each output word. In addition, coverage-penalty is used to favour sentences that cover all the words from the input sentence, with some improvements made to the attention-module [12].

Domain adaptation for NMT

Although the overall results of NMT are nowadays better than those obtained with Statistical Machine Translation (SMT) [13], when a comparative evaluation is performed, NMT systems generate sentences with better fluency, thus sounding more natural, while SMT systems are still better in terms of accuracy [14]. Since NMT uses word embeddings to represent words, this worse accuracy is usually not a big problem provided that the generated words are similar to, or related to, the right word. In the case of the clinical domain, however, accuracy is an important aspect to preserve and some steps must be taken to improve it.

Approaches recently tested with legal domain corpora [15, 16] represent a promising research area in cases in which the sentences from the training corpus are similar. They involve the same basic idea of looking for sentences similar to the input sentence before translating it, but differ in the way this sentence similarity information is used. In one case, this information is used to add the k most similar sentences to the training corpus [15]; in the other, this process is simplified using the sentence similarity scores to rescore the possible output sentences [16].

NMT with low resources

Finally, we refer to the specific task of NMT when limited bilingual resources are available. Taking this problem to the extreme, an emerging interesting research area, known as Unsupervised Machine Translation, is attempting to perform the NMT task without any bilingual data.

Two studies [17, 18] make use of the intrinsic information contained in the word embeddings created from monolingual corpora, and then study the best ways to relate the embedding maps created for each of the languages to be used in the translation process. Both works mark a milestone that changes the traditional paradigm that bilingual corpora is needed to perform NMT, but as expected, they still do not obtain state-of-the-art results compared to NMT systems that make use of bilingual corpora.

There are other well established techniques that help to achieve competitive results with low resources, as in the case of transfer learning [19], based on first training a system with a big enough general corpus and then fine-tuning it with a smaller corpus that can be from a specific domain; or backtranslation [20], based on including a

monolingual corpus and its automatic translation to a bilingual training corpus which is similar in size. Both methods have shown to significantly improve the baseline results when some bilingual data from the domain to be tested is available (in the case of transfer learning), or a monolingual corpus of comparable size to the out-of-domain bilingual corpora is available (for backtranslation).

All these techniques can be beneficial to any language pair in a domain for which there are limited bilingual resources [21].

MATERIALS AND METHODS

System and Equipment

We used the Nematus [9] neural machine translation system, which implements the attention-mechanism on RNNs and makes use of the Theano library, based on Python. Specifically, two different GPU servers were used: one with a Tesla K40 GPU with 12 GB of RAM, and another multi-GPU server with a Titan Xp GPU with 12 GB of RAM.

Resources

Corpora

The out-of-domain corpora used for hyperparameter optimization included a total of 4.5M bilingual sentences. 2.4M of them are a repetition of sentences from the news domain [22], while the remaining 2.2M sentences are from diverse domains such as

administrative, web-crawling and specialised magazines (consumerism and science). These corpora were compiled from diverse sources such as EITB (Basque public broadcaster), Elhuyar (research foundation) and IVAP (official translation service of the Basque Government). Elimination duplication, the effective data expressed in tokens were 102M tokens in Spanish and 72M tokens in Basque.

The Spanish monolingual corpus from the clinical domain was made up of real health records from the hospital of Galdakao-Usansolo consisting of 142,154 documents compiled from 2008 to 2012 with a total of 52M tokens. This dissociated corpus is subject to privacy agreements and is not publicly available.

Table 2 summarises the data of the corpora used.

Domain	Language(s)	Documents	Sentences	Tokens
out-of-domain (news, admin., web-crawling, specialised magazines)	Basque and Spanish	-	4.5M	72M (Basque) 102M (Spanish)
clinical (EHRs)	Spanish	142,154	4.4M	52M

Table 2: Summary of data of the corpora used.

Dictionaries and other resources

Taking advantage of the resources used for the automatic translation of SNOMED CT into Basque [23], a dictionary was built with all the created Basque terms and their corresponding Spanish entries. For many of the Spanish terms refering to a specific SNOMED CT concept, more than one possible Basque term was created. For instance, the Spanish term "lepra" (leprosy) can be translated as "legen", "legen

beltz", "legenar", "negal" or "Hansen-en gaixotasun" (Hansen's disease). In total, the dictionary used for this experiment has 151,111 entries corresponding to 83,360 unique Spanish terms.

Additionally, artificial sentences were created making use of the relations specified on SNOMED CT. Specifically, the Snapshot release of the international version in RF2 format of the SNOMED CT delivery from 2017 July 31st was used. For the sentences to be representative, the most frequent active relations were taken into account, only considering the type of relations that appear more than 10,000 times. The most frequent active relations in the used version were "is a", "finding site", "associated morphology" and "method". As an example, a relation found in the English version is "Uterine hernia" | *is a* | "Disorder of uterus".

Health record templates and manual translations

For evaluating the performance of the system in the clinical domain, a total of 42 health record templates of diverse specializations written in Basque by doctors of the Donostia Hospital [24], and their respective manual translations into Spanish carried out by a bilingual doctor were used as reference. After aligning the sentences obtained from these EHR templates and their respective manual translations, we built a bilingual corpus consisting of 2,076 sentences, that were randomly ordered and divided into 1,038 sentences for development (dev) and 1,038 sentences for testing. Supplementary Table S1 shows the first 10 sentences used for evaluation in the clinical domain.

Our approach

First, we took a model NMT system between Basque and Spanish previously developed using Nematus by one of the authors (G. L.) and performed a hyperparameter optimization based only on the out-of-domain corpus. After this, we progressively added clinical domain resources to measure their influence on translating clinical texts. In this second part, apart from calculating BLEU scores [25], we also carried out a human evaluation by two bilingual doctors who were assisted by professional translators.

NMT hyperparameter optimization

The corpus used for this part of the work was the bilingual out-of-domain one specified in the previous section, with a total of 4.5M sentences. Specifically, 4,530,683 sentences were used for training, 1,994 sentences for development and 1,678 sentences for testing. The latter ones were manually inspected for correctness prior to the testing [10].

The starting point for this part of the work was an NMT system whose basic hyperparameters are shown in Supplementary Table S2.

When choosing the hyperparameters to test, various sources were consulted, but most of the hyperparameters and their possible optimal values were taken from [26]. Supplementary Table S3 shows all the hyperparameters that were tried and their respective values, in the same order in which they were tried.

All the experiments were carried out for both Basque-to-Spanish and Spanish-to-Basque translation directions. After comparing the results for different values of each hyperparameter, the one that achieved the highest BLEU value on the test set was chosen for the next experiment, and, only if the results were significantly different for each translation direction, a different hyperparameter value was selected for each direction.

Evaluation in the clinical domain

First, we chose the system that achieved the best BLEU results on the out-of-domain corpus and subsequently added the following clinical domain resources to the training corpus to measure their incremental contributions to a better translation.

1. A dictionary from the clinical domain

For the first of these experiments, the dictionary used for the automatic translation of SNOMED CT (mentioned in the previous section) was used. For the results to be comparable with the one that only used an out-of-domain corpus, the preprocessing applied after including the dictionary was the same, consisting of tokenization, truecasing and BPE word segmentation.

2. Artificial sentences created from SNOMED CT

For the second of the experiments, artificial sentences created from the relations on SNOMED CT were added. The reason for adding these sentences is that NMT

systems not only learn how to translate words but at the same time learn a language model from the training corpus.

To do so, we first defined two sentence models for each of the most frequent relations in SNOMED CT. Taking these sentence models as a reference, for each of the concepts concerning a unique pair of Basque and Spanish terms, we randomly chose one of the relations that this concept has in SNOMED CT. When doing this, we restricted the possible relations to the most frequent ones and omitted the relations with terms that had not been translated. Finally, we randomly chose one of the sentence models for this specific relation. As an example, figure 1 shows the sentence models in Basque (left) and Spanish (right) corresponding to the previously given "is a" relation example. Equivalent terms are marked with the same color, and the meaning provided by the relation is shown in bold.

uteroaren prolapsoa | umetokiaren nahasmendua | **da**	hernia uterina | **es** | enfermedad uterina
uteroaren prolapsoa | umetokiaren nahasmendua | **mota bat da**	hernia uterina | **es un tipo de** | enfermedad uterina

Figure 1: sentences in Basque (left) and Spanish (right) derived from a relation in SNOMED CT.

QUESTION BY THE ASSOCIATE EDITOR

[What is the meaning of the superscripts 1 and 2 in the figure?? If references for these terms are needed, they should be numbered sequentially with those in the list of References.]

ANSWER FROM THE AUTHORS

(They were used to mark with the same superscript terms that share the meaning. For more clarity, we have changed the figure and the sentence above to mark equivalent terms with the same color.)

Finally, for applying the morphological inflections to the specific terms needed in some of the described sentences in Basque, a transducer was applied following the

inflection rules defined in the Xuxen spelling corrector [27]. After this, a total number of 363,958 sentences were added to the corpus formed by the out-of-domain corpus and the previously added dictionary, carrying out the same preprocessing.

3. A monolingual corpus and its backtranslation

For this part of the work the EHRs from the Spanish monolingual corpus were used. These EHRs were first preprocessed to have one sentence in each line and then the order of the sentences of the set of EHRs was randomly changed to contribute to a better anonymization. For making the translation process faster, repeated sentences were removed from the corpus before translating it, resulting in a total of 2,023,811 sentences that were added to the previous corpus. In order to machine translate these sentences into Basque, the system specified in step 1 was used.

4. A monolingual corpus as bilingual

Finally, following the work described in [28], we also included the same Spanish monolingual corpus and its repetition as if it were Basque, which could be beneficial for the translation of words that do not need to be translated, as in the case of drug names.

These experiments were developed for both Basque-to-Spanish and Spanish-to-Basque translation directions, except for those including the Spanish monolingual corpus, that were performed only for Basque-to-Spanish since the automatically translated corpus can-not be used as a target training corpus [20].

RESULTS

In this section we present the results of our experiments, showing the BLEU values obtained in dev and test sets for the automatic evaluation. Basque-to-Spanish is represented in the tables as "eu-es", while Spanish-to-Basque is represented as "es-eu". As an upper bound reference for BLEU, the state-of-the-art for English-to-German machine translation is 35.0 [29]. The human evaluation is performed in terms of quality and system comparison.

NMT hyperparameter optimization

Table 3 shows the results of the baseline, characterized by the hyperparameter values described in Supplementary Table S2, and the best results obtained with each of the hyperparameters displayed in Supplementary Table S3 for both translation directions in dev and test sets. Note that the results for unit-type correspond to different types Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) for each of the translation directions, as the results indicated this was the best option.

Translation direction	Hyperparameter update	dev BLEU	test BLEU
eu-es	Baseline	26.51	28.98
	Optimizer → Adam	26.87	28.97
	Unit type = GRU	26.87	28.97
	Beam-width → 10	27.21	29.28
	Batch-size → 64	27.02	**29.45**
	Embedding-size → 512	26.65	28.87
es-eu	Baseline	22.95	20.26
	Optimizer → Adam	23.06	20.55
	Unit type → LSTM	23.37	20.96

	Beam-width → 10	23.64	20.93
	Batch-size → 64	23.05	**21.12**
	Embedding-size → 512	23.09	20.42

Table 3: Results for different hyperparameters tested out-of-domain.

Automatic evaluation in the clinical domain

Table 4 shows the BLEU values obtained by adding resources from the clinical domain to the originally out-of-domain training corpus. As all the resources from the clinical domain were added sequentially, the "+" sign should be interpreted as an addition to the corpus corresponding to the immediate upper row.

As stated before, we only tested the inclusion of the Spanish monolingual corpus for Basque-to-Spanish translation direction. We also present the results obtained by Google Translate as a baseline, as this translator will be also taken into account in the human evaluation.

Translation direction	**Training corpus**	**dev BLEU**	**test BLEU**
eu-es	Baseline (Google)	6.16	5.29
	out-of-domain	10.69	10.67
	+ dictionaries	15.45	15.04
	+ artificial sentences	16.08	15.48
	+ backtranslation	22.52	21.07
	+copied	23.57	**21.59**
es-eu	Baseline (Google)	2.28	2.19
	out-of-domain	9.08	8.69
	+ dictionaries	10.75	**10.44**
	+ artificial sentences	10.79	10.43

Table 4: Results in the clinical domain with different training corpora.

Significance

Figure 2 shows the results of applying the Moses script [30] for bootstrap resampling [31] to measure the significance of all the experiments conducted in the clinical domain. To do this, BLEU values are calculated randomly extracting 100 sentences with resampling from the corresponding set, repeating this process 1,000 times, and calculating a confidence interval for the different BLEU values given a p-value (by default, 0.05). As can be observed by comparing the range of BLEU values for each of the systems, only the inclusion of the dictionary and the application of the backtranslation technique for Basque-to-Spanish translation direction gave improvements that could be defined as statistically significant. For both translation directions, the results of Google are significantly lower than the results of any of our systems.

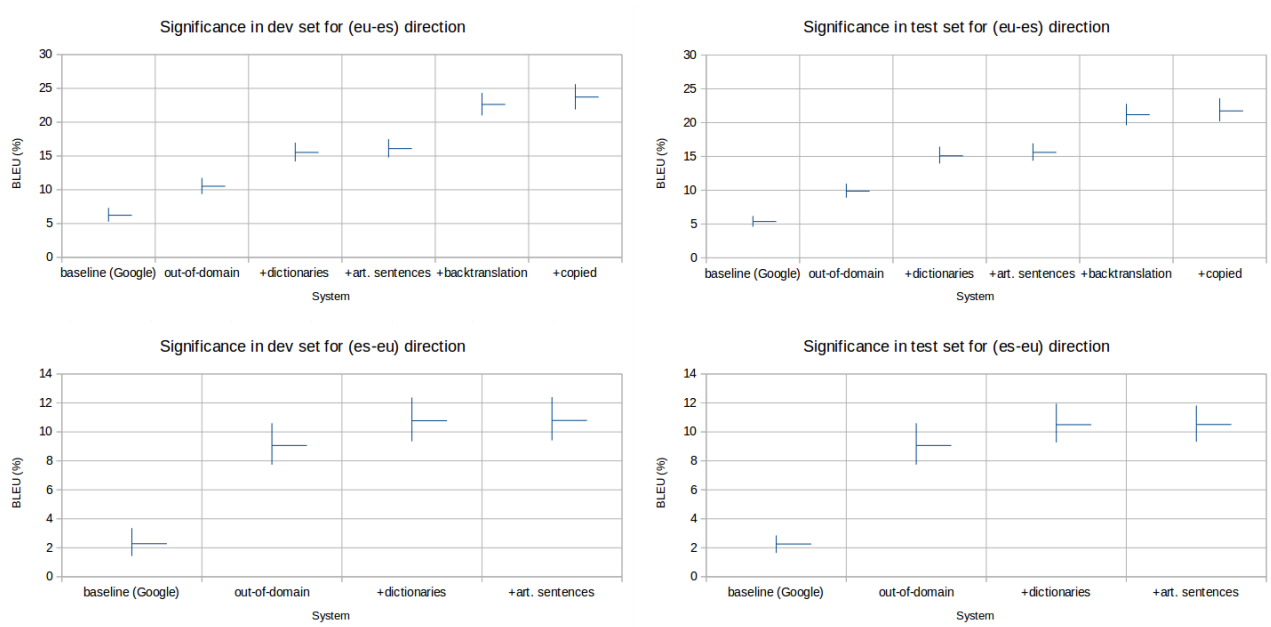

Figure 2: results of applying bootstrap resampling on all the conducted experiments.

Human evaluation in the clinical domain

Here we show the results of the evaluation performed by two bilingual doctors using the translation evaluation tool provided by TAUS [32] on 100 non-repeated sentences randomly extracted from the test set. We carried out two kinds of evaluation, one for ranking the different translations made by 1) a human (reference used for the automatic evaluation), 2) the IxaMedNMT-RNN system (our best performing system in the clinical domain) and 3) Google Translate (baseline used in the previous section); and another for evaluating our IxaMedNMT-RNN system, obtaining fluency and adequacy scores (from 1 to 4), as well as the number of fluency, accuracy, terminology, style and locale convention errors in each sentence.

Figure 3 shows the comparison of the rankings given by each evaluator to the translations of the different systems. Both evaluators generally agree that the human translator is slightly better than the IxaMedNMT-RNN system, while this is much better than Google Translate. We calculated Cohen's kappa for measuring inter-annotator agreement, obtaining a 0.25 value for Human Vs IxaMedNMT-RNN comparison (fair agreement) and 0.17 for IxaMedNMT-RNN Vs Google comparison (slight agreement).

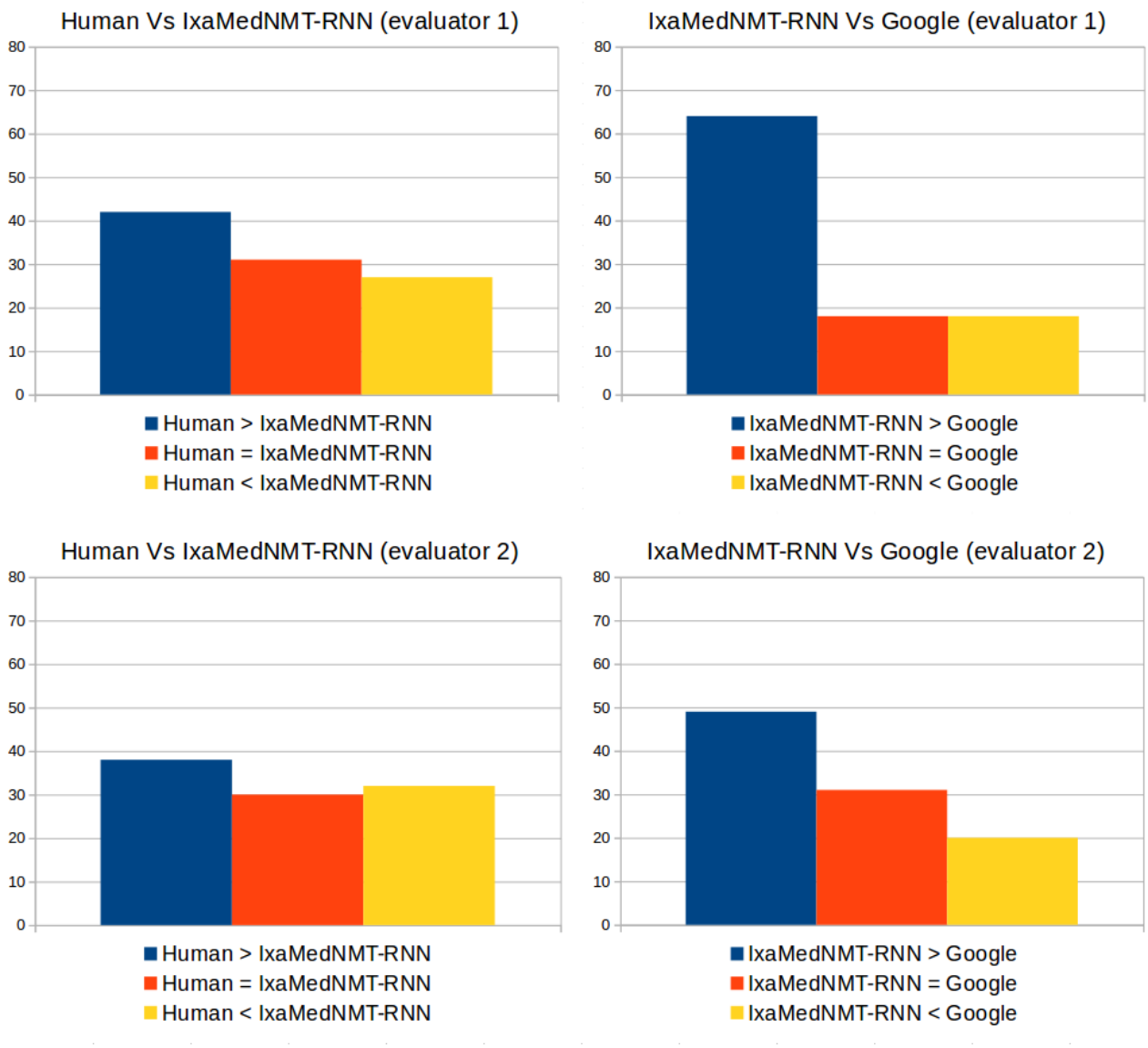

Figure 3: comparison between Human Vs IxaMedNMT-RNN (left) and IxaMedNMT-RNN Vs Google (right) scores given by human evaluators.

In figure 4 we provide the fluency and adequacy scores (top), together with fluency and accuracy errors (bottom) given by each evaluator. We observe that more sentences are ranked as flawless (score: 4) in terms of fluency than adequacy, while the number of accuracy errors seems to be more distributed among the translated sentences. The kappa coefficients are 0.15 (slight agreement) for fluency score and 0.65 (substantial agreement) for adequacy score. Note that for this figure and the next one, we omit the out-of-range number of errors corresponding to a sentence

containing only medical analysis results, which got 20 accuracy errors according to evaluator 1 and 14 terminology errors according to evaluator 2.

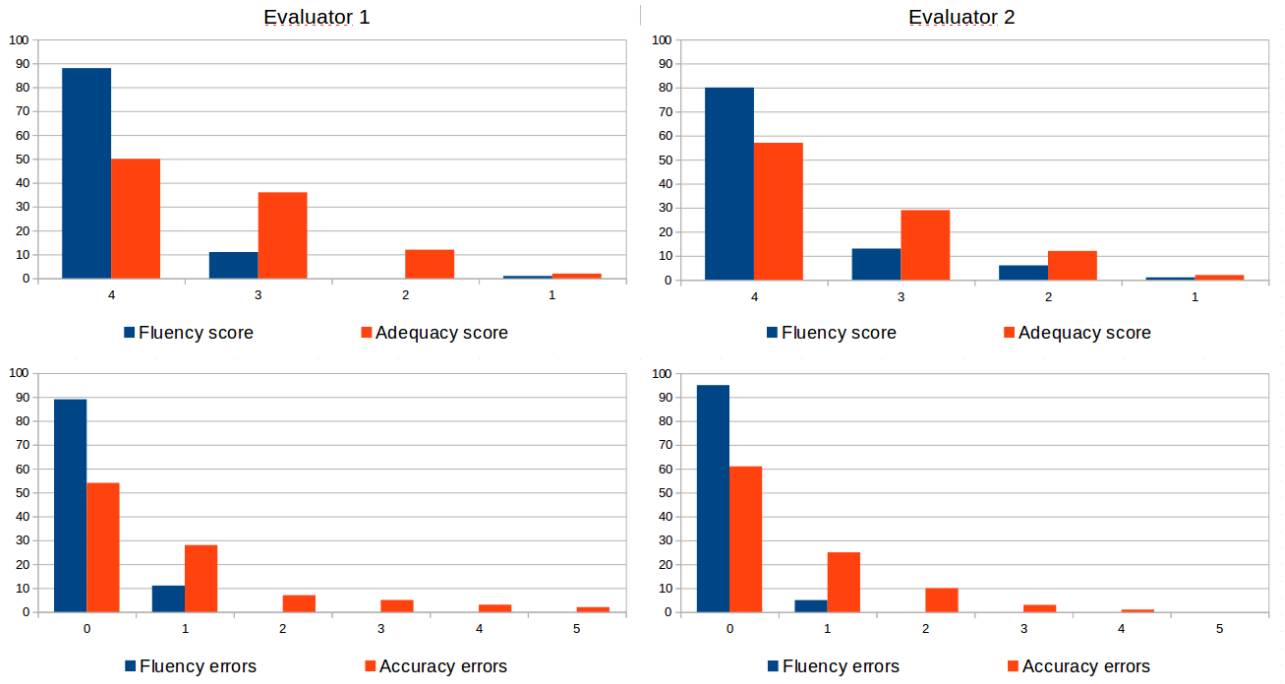

Figure 4: Fluency and adequacy scores (top); number of fluency and accuracy errors (bottom).

Finally, figure 5 shows the number of terminology and locale convention errors. Evaluator 1 detected 3 sentences with 1 terminology error, while evaluator 2 marked 7 sentences with 1 terminology error and 2 with 2 errors. For locale convention, evaluator 1 detected 2 separated errors, while evaluator 2 only marked one of them, being the other one a date kept in Basque format (*yyyy-mm-dd*). None of the evaluators detected any style error in the tested sentences.

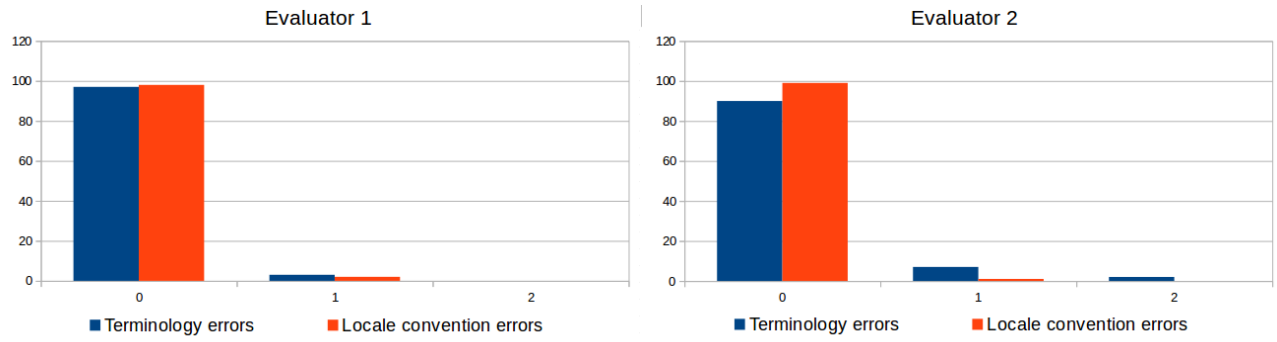

Figure 5: number of terminology and locale convention errors.

Translation example

Figure 6 shows a clinical domain translation example.

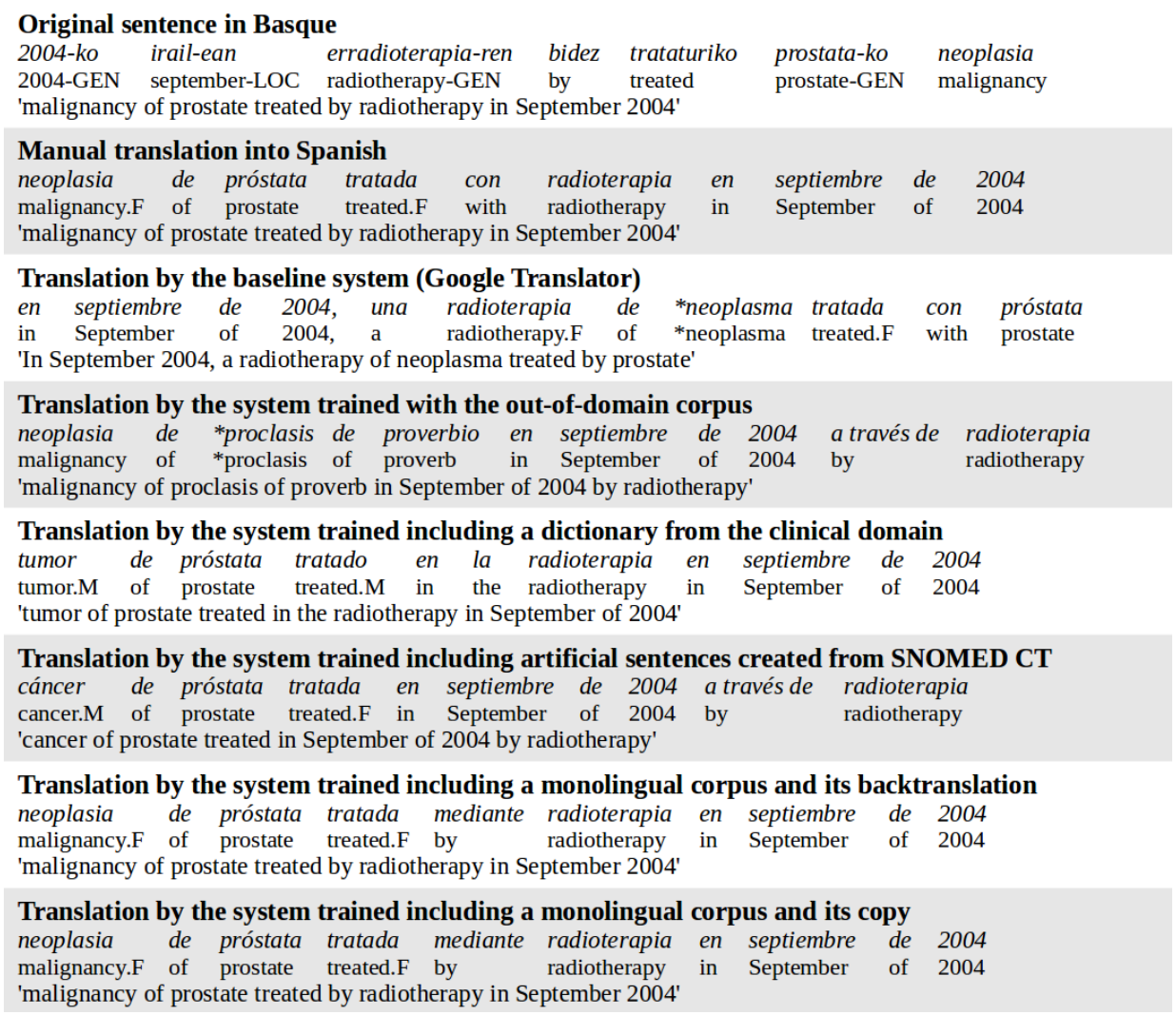

Figure 6: example of a sentence translated by the different systems tested in the clinical domain.

The Google translation, although containing almost all the correct terms, loses the original meaning. Our system trained out-of-domain is unable to translate "prostata", and misses the term "tratada" (treated). The systems including the dictionary and the artificial sentences are incapable of reproducing "neoplasia", giving the inexact

"tumor" and "cancer". Finally, the systems leveraging the clinical domain monolingual corpus produce flawless and adequate translations.

DISCUSSION

Basque being a morphologically rich language, and having used the BLEU metric that counts the number of words and n-grams correctly translated, higher values are expected for Basque-to-Spanish than for Spanish-to-Basque.

When analyzing the results of hyperparameter optimization (Table 3), we observe a 0.47 points increase in the test set for Basque-to-Spanish; while for Spanish-to-Basque the improvement is of 0.86 points. In the case of Basque-to-Spanish, the improvement came from changing the values of beam-width and batch-size, while for Spanish-to-Basque the results improved when changing the optimizer, unit-type and batch-size.

Therefore, we can conclude that the conducted experiments were mostly satisfactory (except for the embedding-size) and further experiments should be carried out for both beam-width and batch-size.

Analyzing the results in the clinical domain (Table 4), it can be noted that all the conducted experiments improved the results, except for the inclusion of artificial sentences that proved to be non-beneficial, especially for Spanish-to-Basque. We believe that this happened because the sentence models based on SNOMED CT relations were very simple and their syntax was already represented in the out-of-domain corpus, while the terminology was included in the dictionaries.

Regarding the different translation directions, it can be seen that the inclusion of each of the resources from the clinical domain has been more useful for Basque-to-Spanish.

We highlight the inclusion of the dictionary, where a 4.4 BLEU points gain was achieved in the test set for Basque-to-Spanish, compared to a 1.7 points increase for Spanish-to-Basque. Given the existence of translations of SNOMED CT into many languages, a similar dictionary resource might be generated for other language pairs for which bilingual clinical corpora are lacking.

Finally, examining the results of including the different resources from the clinical domain, we conclude that the inclusion of the Spanish monolingual corpus and its translation into Basque has been the most beneficial, followed by the inclusion of the dictionary. Both results reflect that health records make use of a very specific vocabulary and syntax, which is shown by these great improvements with the inclusion of a relatively small dictionary and a synthetic bilingual corpus formed by a monolingual corpus and its machine translation. We demonstrate that the backtranslation technique, while simple, is highly effective because it helps the decoder to perform the language modeling task better.

The human evaluation confirmed these good results, ranking our system much closer to the human reference translation than to the automatic baseline system; and achieving high fluency and adequacy scores for most of the tested sentences.

For future experiments, we have to point out that even if bilingual corpora from the clinical domain becomes available, the application of the backtranslation technique

will also be helpful, as most of the state-of-the-art systems make use of this technique to improve their results.

CONCLUSION

We managed to optimize NMT hyperparameter values on an out-of-domain corpus, with almost 0.5 points gain in BLEU for Basque-to-Spanish, and almost 0.9 points improvement for Spanish-to-Basque from an already strong baseline.

Regarding the evaluation in the clinical domain, we point out the great improvement achieved through the technique of backtranslation, with a 5.6 BLEU points gain for the tested Basque-to-Spanish translation direction. We also observe that the inclusion of the dictionary from the clinical domain has significantly improved the results, especially for Basque-to-Spanish, obtaining a 4.4 BLEU points gain. Altogether, the applied improvements have made it possible to approach the out-of-domain results, raising an acceptable result of 21.59 BLEU points for Basque-to-Spanish. These automatic evaluation results were confirmed by the human evaluation performed, showing that it is possible to develop a NMT system useful for translating clinical texts without making use of any bilingual corpus from the clinical domain.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We would like to thank Nora Aranberri, Idoia Arrizabalaga, Natalia Elvira and Aitziber Etxagibel for helping us to perform the human evaluation. We would also like to thank Uxoa Iñurrieta for helping us with the glosses included in the translation example.

FUNDING STATEMENT

This work was supported by the Spanish Ministry of Economy and Competitiveness grant numbers BES-2017-081045, TIN2015-70214-P (TADEEP) and TIN2016-77820-C3-1-R (PROSA-MED).

COMPETING INTERESTS STATEMENT

The authors have no competing interests to declare.

CONTRIBUTORSHIP STATEMENT

All authors have made substantial contributions to the conception or design of the work. Xabier Soto was responsible for drafting the work, while Olatz Perez-de-Viñaspre, Gorka Labaka and Maite Oronoz critically revised it for important intellectual content. All authors have given final approval to the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

REFERENCES

1 Google Translate. https://translate.google.com/. Accessed May 20, 2019.

2 Forcada ML, Ñeco RP. Recursive hetero-associative memories for translation. In International Work-Conference on Artificial Neural Networks. Springer, 1997:453–462.

3 Castaño A, Casacuberta F. A connectionist approach to machine translation. In Fifth European Conference on Speech Communication and Technology, 1997.

4 Kalchbrenner N, Blunsom P. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013:1700–1709.

5 Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, 2014:3104–3112.

6 Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature*, 1986;323(6088):533.

7 Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

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

9 Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909, 2015.

10 Etchegoyhen T, Martínez E, Azpeitia A, et al. Neural machine translation of basque. In Proceedings of the 21st Annual Conference of the European Association for Machine Translation, 2018:139–148.

11 Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.

12 Tu, Z, Lu, Z, Liu, Y, et al. Modeling coverage for neural machine translation. arXiv preprint arXiv:1601.04811, 2016.

13 Bojar O, Chatterjee R, Federmann C, et al. Findings of the 2016 conference on machine translation. In ACL 2016 FIRST CONFERENCE ON MACHINE

TRANSLATION (WMT16). The Association for Computational Linguistics, 2016:131–198.

14 Koehn P, Knowles R. Six challenges for neural machine translation. arXiv preprint arXiv:1706.03872, 2017.

15 Gu J, Wang Y, Cho K, et al. Search engine guided non-parametric neural machine translation. arXiv preprint arXiv:1705.07267, 2017.

16 Zhang J, Utiyama M, Sumita E, et al. Guiding neural machine translation with retrieved translation pieces. arXiv preprint arXiv:1804.02559, 2018.

17 Artetxe M, Labaka G, Agirre E, et al. Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041, 2017.

18 Lample G, Denoyer L, Ranzato MA. Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv:1711.00043, 2017.

19 Zoph B, Yuret D, May J, et al. Transfer learning for low-resource neural machine translation. arXiv preprint arXiv:1604.02201, 2016.

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

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

22 Etchegoyhen T, Azpeitia A, Perez N. Exploiting a Large Strongly Comparable Corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2016.

23 Perez-de-Viñaspre O. Automatic medical term generation for a low-resource language: translation of SNOMED CT into Basque. PhD thesis, University of the Basque Country, Donostia, Euskal Herria, 2017.

24 Joanes Etxeberri Saria V. Edizioa. Donostia Unibertsitate Ospitaleko alta-txostenak. Donostiako Unibertsitate Ospitalea, Komunikazio Unitatea, 2014.

25 Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002:311–318.

26 Britz D, Goldie A, Luong T, et al. Massive exploration of neural machine translation architectures. arXiv preprint arXiv:1703.03906, 2017.

27 Agirre E, Alegria I, Arregi X, et al. Xuxen: A spelling checker/corrector for basque based on two-level morphology. In Proceedings of the third conference on Applied natural language processing. Association for Computational Linguistics, 1992:119-125.

28 Currey A, Barone AVM, Heafield K. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In Proceedings of the Conference on Machine Translation (WMT). Association for Computational Linguistics, 2017;1:148-156.

29 Edunov S, Ott M, Auli M et al. Understanding back-translation at scale. arXiv preprint arXiv:1808.09381, 2018.

30 Script used for measuring the significance by bootstrap resampling. https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/bsbleu.py. Accessed May 20, 2019.

31 Koehn P. Statistical significance tests for machine translation evaluation. In Proceedings of the 2004 conference on empirical methods in natural language processing 2004.

32 Webpage of the evaluation tool used for the human evaluation. https://taus.net/dqf/. Accessed May 20, 2019.