Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish

Xabier Soto, Olatz Perez-de-Viñaspre, Maite Oronoz, Gorka Labaka
Ixa Research Group, University of the Basque Country (UPV/EHU)
{xabier.soto, olatz.perezdevinaspre, maite.oronoz, gorka.labaka}@ehu.eus

Abstract

We present a method for machine translation of clinical texts without using bilingual clinical texts, leveraging the rich terminology and structure of the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), which is considered the most comprehensive, multilingual clinical health care terminology collection in the world. We evaluate our method for Basque to Spanish translation, comparing the performance with and without using clinical domain resources. As a method to leverage domain-specific knowledge, we incorporate to the training corpus lexical bilingual resources previously used for the automatic translation of SNOMED CT into Basque, as well as artificial sentences created making use of the relations specified in SNOMED CT. Furthermore, we use available Electronic Health Records in Spanish for backtranslation and copying. For assessing our proposal, we use Recurrent Neural Network and Transformer architectures, and we try diverse techniques for backtranslation, using not only Neural Machine Translation but also Rule-Based and Statistical Machine Translation systems. We observe large and consistent improvements ranging from 10 to 15 BLEU points, obtaining the best automatic evaluation results using Transformer for both general architecture and backtranslation systems.

1 Introduction

The objective of this work is to study the utility of available clinical domain resources in a real use-case, which is the translation of Electronic Health Records (EHR) from Basque to Spanish. Basque is a minoritised language, also in the Basque public health service, where most of the EHRs are written in Spanish so that any doctor can understand them. With the aim of enabling Basque speaking doctors to write EHRs in Basque, we have the long-term objective of developing machine translation systems to translate clinical texts between Basque and Spanish. This work presents a method for machine translation of clinical texts from Basque to Spanish, conditioned by the current lack of clinical domain corpora in Basque.

Neural Machine Translation (NMT) has become in the past recent years the prevailing technology for machine translation, especially in the research community. Several architectures have been proposed for NMT, ranging from the initial Convolutional Neural Networks (CNN) (Kalchbrenner and Blunsom, 2013) and Recurrent Neural Networks (RNN) (Sutskever et al., 2014), to the most advanced Transformer (Vaswani et al., 2017). However, it is known that NMT systems require a large amount of training data to obtain optimal results (Koehn and Knowles, 2017), so traditional techniques as Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT) (Koehn et al., 2003) can be considered when the available resources are low.

One of the techniques that has become a standard to increase the available resources for NMT systems is backtranslation (Sennrich et al., 2015a), consisting in automatically translating a monolingual corpus from the target language into the

source language, and then adding both original and translated corpora to the training corpus. In our case, the availability of EHRs in Spanish enables us to improve the results for the translation of clinical texts from Basque to Spanish, also serving us as a resource for domain adaptation.

Another of our challenges is to study how to translate clinical text, which has its own characteristics differentiated from texts from other domains. Usually, the grammar of the sentences in EHRs is simplified, often omitting verbs, missing punctuation, using many acronyms and with a non-standard language more oriented to communicate between doctors than for being understood by patients. Furthermore, the main difficulty of translating clinical texts comes from the rich vocabulary used in EHRs to refer to drugs, diseases, body parts and other clinical terminology.

Regarding the language pair, our main challenge is to deal with long distance languages as Basque and Spanish, with the complexity associated with it. Specifically, we have to address the challenge of translating from a language with the characteristics of Basque. Briefly, Basque language can be described as a highly agglutinative language, with a rich morphology, where words are usually created adding diverse suffixes that mark different cases. The morphology of verbs is especially complex, including morphemes that add information about the subject, object, number, tense, aspect, etc. It is thought that the BPE word segmentation commonly used in NMT (Sennrich et al., 2015b), originally developed for avoiding the out-of-vocabulary problem, is also beneficial for the translation from morphologically rich languages as Basque.

2 Related work

Several approaches have been tried for machine translation of Basque, including Example-Based (Stroppa et al., 2006), Rule-Based (Mayor, 2007) and Statistical systems (Labaka, 2010). First works have been published for Neural Machine Translation of Basque (Etchegoyhen et al., 2018; Jauregi et al., 2018), and the first general domain commercial system for NMT between Basque and Spanish is already available online.[1]

In the NMT approach for Basque by Etchegoyhen et al. (2018), diverse morphological segmentation techniques are tested, including the aforementioned Byte Pair Encoding (BPE) (Sennrich et al., 2015b), the linguistically motivated vocabulary reduction originally proposed for Turkish (Ataman et al., 2017) and the ixaKat morphological analyser for Basque (Alegria et al., 1996; Otegi et al., 2016). They also tried character-based Machine Translation (Lee et al., 2016), obtaining the best results for translating from Basque to Spanish when applying the morphological analyser for Basque followed by BPE word segmentation to the source language corpus, and only BPE word segmentation to the target language corpus.

Regarding the clinical domain, Perez-de-Vinaspre (2017) developed a system for automatically translating the clinical terminology included in SNOMED CT (IHTSDO, 2014) into Basque. Perez-de-Vinaspre (2017) combined the use of lexical resources, transliteration of neoclassic terms, generation of nested terms and the adaptation of a RBMT system for the medical domain as backup. With respect to the translation of EHRs, the bibliography is scarce, and nowadays we can only refer to a preliminary study for translating clinical notes from English to Spanish (Liu and Cai, 2015).

Another approach for the task of translation of clinical texts is domain adaptation. Usually, when low resources for the desired domain are available, a bigger corpus from another domain is used to first train the system, which is then fine-tuned with the available in-domain corpus (Zoph et al., 2016). From another point of view, Bapna and Firat (2019) try to combine non-parametric or retrieval based approaches with NMT, looking for similarities between n-grams in the sentence to be translated and part of previously translated sentences, and then using this information for producing more accurate translations.

Concerning backtranslation, we have considered the analysis performed by Poncelas et al. (2018), where different sizes of backtranslated corpora were added to the human translated corpora used as training corpus; and regarding the techniques used for backtranslation, we follow the work by Burlot and Yvon (2019) in which they compare the performance of different SMT and NMT systems for this task.

3 Resources and methodology

As mentioned in the introduction, our main handicap is the lack of clinical domain bilingual corpora. To overcome this, we make use of available out-

[1] https://www.modela.eus/eu/itzultzailea (Accessed on April 11, 2019.)

of-domain bilingual corpora, automatically created clinical terminology in Basque (Perez-de-Viñaspre, 2017), artificial sentences formed based on the relations specified in SNOMED CT, and EHRs in Spanish that are used for backtranslation (Sennrich et al., 2015a) and copying (Currey et al., 2017).

For evaluation in the clinical domain, we use EHR templates in Basque published with academic purposes (Joanes Etxeberri Saria V. Edizioa, 2014), together with their manual translations into Spanish performed by a bilingual doctor.

In the following, we present the details of each of the resources and explain how they were used in this work.

3.1 Out-of-domain corpora

As a basis for our work, we use a large bilingual corpus formed by 4.5 million sentences, where 2.3 million sentences are a repetition of a corpus from the news domain (Etchegoyhen et al., 2016), and the remaining 2.2 million sentences are from diverse domains such as administrative, web-crawling and specialised magazines (consumerism and science). These corpora were compiled from sources such as EITB (Basque public broadcaster), Elhuyar (research foundation) and IVAP (Basque institute of public administration).

3.2 Clinical terminology

As a first step for improving the translation of clinical texts, we built a dictionary with all the Basque terms and their corresponding Spanish entries used for the automatic translation of SNOMED CT into Basque (Perez-de-Viñaspre, 2017). These terms were compiled from different sources such as Euskalterm, Elhuyar Science and Technology dictionary, UPV/EHU human anatomy atlas and nursery dictionary, International Classification of Diseases dictionary and a health administration related dictionary. As this work corresponds to a first approach of developing a Basque version of SNOMED CT, more than a possible Basque term was created for each entry in Spanish. Altogether, we use 151,111 Basque terms corresponding to 83,360 unique Spanish terms. We think that the fact of having more than one possible Basque term for each Spanish entry helps us to improve the coverage of the system for translating from Basque to Spanish. As a sample, Table 1 shows the first 10 clinical terms included as training corpus.

3.3 Artificial sentences

While including clinical terms in our system helps us to approach the rich terminology characteristic of clinical notes, we think that including these same terms in the form of sentences could be more suitable to the task of translating sentences from EHRs. For doing this, we leverage the structured form of SNOMED CT, using the relations specified in it to create simple artificial sentences that could be more similar to the real sentences included in EHRs.

Specifically, the Snapshot release of the international version on RF2 format of the SNOMED CT delivery from 31st July 2017 was used. For the sentences to be representative, the most frequent active relations were taken into account, only considering the type of relations that appeared more than 10,000 times. The most frequent active relations in the used version were "is a", "finding site", "associated morphology" and "method".

For creating the artificial sentences, we first defined two sentence models for each of the most

Basque term	**Spanish term**	**English gloss**
organo kopulatzaile	órgano copulador	*copulatory organ*
dionisiako	dionisiaco	*Dionysian*
desfile	desfile	*parade*
begi-miiasia	miasis ocular	*ophthalmic myiasis*
ahoko kandidiasi	candidiasis oral	*oral candidiasis*
wolfram	wolframio	*Tungsten*
W	wolframio	*Tungsten*
zergari	recaudador	*collector*
jasotzaile	recaudador	*collector*
biltzaile	recaudador	*collector*

Table 1: First 10 clinical terms included as training corpus.

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 were not available in both languages. Finally, we randomly chose one of the two sentence models for this specific relation.

Considering the agglutinative character of Basque language, some of the created sentences needed the application of morphological inflections to the specific terms included in the artificial sentences. For this task, a transducer was applied following the inflection rules defined in the Xuxen spelling corrector (Agirre et al., 1992). In total, 363,958 sentences were created. As a sample, Table 2 shows the first 10 artificial sentences created with this method, separating different terms and relations with '|', giving the same superscript number to equivalent terms, and marking the terms that define the relations in bold.

3.4 EHRs in Spanish

Finally, as a main contribution to the translation of clinical texts, we make use of available EHRs in Spanish. This corpus is made up of real health records from the hospital of Galdakao-Usansolo consisting of 142,154 documents compiled from

Basque sentence	**Spanish sentence**
umetokiaren prolapsoa[1] | emakumezkoaren prolapso genitala, zehaztugabea[2] | **da** *uterine prolapse*[1] | ***is a*** | *prolapse of female genital organs, undefined*[2]	prolapso uterino[1] | **es** | prolapso de los órganos genitales femeninos[2] *uterine prolapse*[1] | ***is a*** | *prolapse of female genital organs*[2]
umetokiaren prolapsoa[1] | uteroa[2] | **-n gertatzen da** *uterine prolapse*[1] | ***occurs in*** | *uterus*[2]	descenso uterino[1] | **ocurre en** | estructura uterina[2] *descensus uteri*[1] | ***occurs in*** | *uterine structure*[2]
umetokiaren prolapsoa[1] | uteroaren egitura[2] | **-n aurkitzen da** *uterine prolapse*[1] | ***is found in*** | *uterine structure*[2]	hernia uterina[1] | **se encuentra en** | estructura uterina[2] *uterine hernia*[1] | ***is found in*** | *uterine structure*[2]
uteroaren prolapsoa[1] | emakumezkoaren prolapso genitala, zehaztugabea[2] | **da** *uterine prolapse*[1] | ***is a*** | *prolapse of female genital organs, undefined*[2]	prolapso uterino[1] | **es** | prolapso genital[2] *uterine prolapse*[1] | ***is a*** | *genital prolapse*[2]
uteroaren prolapsoa[1] | umetokiko trastorno ez-inflamatorioa, zehaztugabea[2] | **mota bat da** *uterine prolapse*[1] | ***is a type of*** | *noninflammatory uterine disorder, undefined*[2]	descenso uterino[1] | **es un tipo de** | trastorno uterino[2] *descensus uteri*[1] | ***is a type of*** | *uterine disorder*[2]
uteroaren prolapsoa[1] | umetokiaren nahasmendua[2] | **da** *uterine prolapse*[1] | ***is a*** | *uterine disorder*[2]	hernia uterina[1] | **es** | enfermedad uterina[2] *uterine hernia*[1] | ***is a*** | *uterine disease*[2]
zakilaren inflamazioa[1] | zakil[2] | **-ean gertatzen da** *inflammation of penis*[1] | ***occurs in*** | *penis*[2]	inflamación del pene[1] | **ocurre en** | estructura de pene[2] *inflammation of penis*[1] | ***occurs in*** | *penis structure*[2]
zakilaren inflamazioa[1] | zakilaren egitura[2] | **-n aurkitzen da** *inflammation of penis*[1] | ***is found in*** | *penis structure*[2]	trastorno inflamatorio del pene[1] | **se encuentra en** | pene[2] *inflammatory disorder of penis*[1] | ***is found in*** | *penis*[2]
zakilaren hantura[1] | zakilaren gaitza[2] | **da** *inflammation of penis*[1] | ***is a*** | *disorder of penis*[2]	inflamación del pene[1] | **es** | enfermedad peniana[2] *inflammation of penis*[1] | ***is a*** | *disorder of penis*[2]
zakilaren hantura[1] | zakilaren gaitz[2] | **mota bat da** *inflammation of penis*[1] | ***is a type of*** | *disorder of penis*[2]	trastorno inflamatorio del pene[1] | **es un tipo de** | enfermedad peniana[2] *inflammatory disorder of penis*[1] | ***is a type of*** | *disorder of penis*[2]

Table 2: First 10 artificial sentences created from relations in SNOMED CT.

2008 to 2012. Due to privacy agreements, this dissociated corpus is not publicly available.

These documents were first preprocessed to have one sentence in each line, and then the order of the sentences was randomly changed to contribute to a better anonymisation. 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.

This corpus was added twice to the training corpus, once by applying different backtranslation techniques, and the other by simply using the same corpus in Spanish as if it were Basque (Currey et al., 2017), which we think could be beneficial for the translation of words that do not need to be translated, as it is the case of drug names. This way, from the total number of sentences used for training the corpus based systems developed for translation of clinical texts (9,093,374), around half of them correspond to out-of-domain sentences (4,530,683), and the other half come from diverse clinical domain sources (4,562,691).

Table 3 summarises the numbers of the training corpora. All corpora was tokenised and truecased using the utilities of Nematus (Sennrich et al., 2017) if they were to be used for corpus based systems. For NMT experiments, BPE word segmentation was performed using subword-nmt[2], applying 90,000 merge operations on the joint bilingual corpora. The number of tokens in Basque for the backtranslated EHRs correspond to the backtranslation performed with the shallow RNN.

3.5 EHR templates in Basque and their manual translations into Spanish

For evaluating the task of translating clinical texts, we used 42 EHR templates of diverse specializations written in Basque by doctors of the Donostia Hospital, and their respective manual translations into Spanish carried out by a bilingual doctor. We manually aligned the sentences from these templates with their respective translations, building a bilingual corpus of 2,076 sentences. These sentences were randomly ordered and further divided into 1,038 sentences for development purposes and 1,038 sentences for test purposes.

We highlight that the sentences used for evaluation in the clinical domain come from diverse specializations, which we expect to be mirrored in a more diverse set of development and test corpora. Furthermore, from the 1,038 sentences from the test set, 826 are non-repeated, corresponding the most repeated ones to short sentences relating to EHR section titles. As a sample, Table 4 shows the first 10 sentences used for evaluation in the clinical domain.

4 Experiments

We test our method through two types of experiments, one regarding different NMT architectures, and the other referring to different systems used for backtranslation. All the experiments concerning NMT systems were performed on Titan XP GPUs, using only one for training the shallow RNN, and two for the deep RNN and the Transformer.

4.1 Architectures

First, we test the performance of several neural architectures, trying a shallow RNN as an easily reproducible system, a Transformer (Vaswani et al., 2017) architecture as state-of-the-art performing system, and a deep RNN (Barone et al., 2017) as a fairer comparison to Transformer.

We develop two systems for each architecture, one trained only with out-of-domain corpora, and another trained with all the available resources, including the ones from the clinical domain. For this part of the work, the backtranslation of the available EHRs in Spanish was performed by the shallow RNN.

We evaluate the performance of all the systems in the clinical domain, using the EHR templates in

[2] https://github.com/rsennrich/subword-nmt (Accessed on April 11, 2019.)

Domain	Type	Sentences	Tokens
out-of-domain	Diverse sentences	4.5 million	73 million (Basque) / 102 million (Spanish)
clinical domain	Terms	151,111	271,248 (Basque) / 257,641 (Spanish)
	Artificial sentences	363,958	3.1 million (Basque) / 4.1 million (Spanish)
	Backtranslated EHRs	2 million	26 million (Basque) / 33 million (Spanish)
	Copied EHRs	2 million	33 million

Table 3: Numbers of the training corpora.

Basque sentence	Spanish sentence
tratamendua *therapy*	tratamiento *therapy*
abortuak: 1 *aborta 1*	abortos 1 *aborta 1*
lehenengo sintomatologia *first symptomatology*	primera sintomatología *first symptomatology*
fibrinolisiaren ondoren egoera klinikoa ez da askorik aldatu *clinical status does not change much after fibrinolysis*	la situación clínica después de la fibrinólisis no cambia sustancialmente *clinical status after fibrinolysis does not change substantially*
hipertentsioaren aurkako tratamenduarekin hasi da, tentsioak neurri egokian mantenduz; hipergluzemiarako joera antzeman da egonaldian *he/she started the treatment for hypertension, keeping tensions at the right level; a tendency to hyperglycemia is observed during the stay*	al mismo tiempo tratamientopara normalizar la HTA, hiperglucemia y dislipemia *at the same time treatmentfor* normalising HBP, hyperglycemia and dislipemia**
ebakuntza aurreko azterketa normala izan ostean, 2012-08-20an operazioa egin da *after the preoperative examination being normal, the operation is done on 2012-08-20*	tras ser normal la exploración preoperatoria se opera el 20-08-2012, practicándose: *after the preoperative exploration being normal he/she is operated on 2012-08-20, by practising:*
Dismetriarik ez *No dysmetria*	no dismetría *no dysmetria*
miaketa oftalmologikoa normala *normal ophthalmic exploration*	examen oftalmológico normal *normal ophthalmic examination*
EKG: erritmo sinusala, 103 tau/min *ECG: sinus rhythm, 103 beat/min*	EKG-ritmo sinusal 103/minuto *ECG-sinus rhythm, 103/min*
ez du botaka egin *he/she has not vomited*	no vómitos *no vomits*

Table 4: First 10 sentences used for evaluation in the clinical domain.

Basque and their manual translations into Spanish specified in the previous section.

A description of the tested architectures is given in the following lines.

Shallow RNN: As a simple RNN, we use a model developed with the old version of Nematus (Sennrich et al., 2017), making use of the Theano framework. Specifically, we use 1 layer (bidirectional for the encoder) of 1024 GRU (Cho et al., 2014) units, with a embedding-size of 500, a batch-size of 64 and using Adam (Kingma and Ba, 2014) as optimisation method. For decoding, we use a beam-width of 10 for all the experiments. Some of the values of these hyperparameters were optimised with the out-of-domain corpus, and subsequently used in the other architectures.

Deep RNN: As a more advanced RNN, we select the system developed by Barone et al. (2017), included in a more recent work in which linguistic abilities of diverse NMT systems were tested (Tang et al., 2018).

From the different variants presented in Barone et al. (2017), we use the one that obtained the best reported results, whose configuration parameters are public.[3]

Transformer: As a state-of-the-art NMT system, we choose the Transformer implementation in Pytorch by OpenNMT (Klein et al., 2017). We use the recommended hyperparameters,[4] except for the number of GPUs and batch-size, that were

[3] https://github.com/Avmb/deep-nmt-architectures/blob/master/configs/bideep-bideep-rGRU-large/config.sh (Accessed on April 11, 2019.)

[4] http://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model-do-you-support-multi-gpu (Accessed on April 11, 2019.)

halved to meet our hardware capabilities.

4.2 Backtranslation systems

After trying different architectures, we select the one that obtains the best automatic evaluation results in the clinical domain and change the way the backtranslation is performed. For that, we compare the shallow RNN architecture with the one that gets the best results in the clinical domain, and also try RBMT and SMT systems to translate the EHRs in Spanish into Basque.

For training the corpus based systems in the Spanish-to-Basque translation direction, we use the out-of-domain corpus and the dictionaries including clinical terminology. The resulting synthetic corpus is added together with the artificial sentences and the copied monolingual corpus, and the performance of the systems is tested in the clinical domain.

Shallow RNN: For this experiment we use the same shallow RNN architecture specified in the previous section, just changing the translation direction. Note that, due to an error in the preprocessing, the BPE word segmentation was performed for 45,000 steps in each language corpus, instead of 90,000 times in the joint corpora. We do not expect for this error to have significant influence on the final results.

Transformer: We train the Transformer system in the Spanish-to-Basque translation direction with the same hyperparameters specified in the previous section. Following the work by Edunov et al. (2018), we perform the translation by unrestricted random sampling, which is proved to obtain better results than restricted random sampling or traditional beam search when applied to backtranslation.

RBMT: For this part of the work, we try Matxin (Mayor, 2007), a Rule-Based system for Spanish-to-Basque Machine Translation, adapted to the biomedical domain by the inclusion of dictionaries. In this case, we translate the EHRs in Spanish before truecasing, so when removing the repeated sentences from the corpora the number of sentences is not exactly the same as for the monolingual corpus translated with corpus based systems (2,036,165 instead of 2,023,811).

SMT: Finally, we try Moses (Koehn et al., 2007) as a statistical system, adapted to the biomedical domain. We use default parametrisation with MGIZA for word alignment, a "msd-bidirectional-fe" lexicalised reordering model and a KenLM (Heafield, 2011) 5-gram target language model. The weights for the different components were adjusted to optimise BLEU using Minimum Error Rate Training (MERT) with an n-best list of size 100.

5 Results and discussion

In this section we show and discuss the automatic evaluation results of the experiments carried out with different architectures and backtranslation systems. In both cases, we calculate BLEU (Papineni et al., 2002) in development and test sets using the multi-bleu script included in Moses.[5]

5.1 Architectures

Table 5 shows the results of the tested architectures in two variants: 1) trained only with out-of-domain corpora, and 2) including all the clinical domain resources. We observe large and consistent improvements when adding in-domain data to each of the tested architectures. Surprisingly, the deep RNN obtains lower results than the shallow RNN, especially comparing the systems trained out-of-domain, which can be an overfitting issue. We also think that the previous optimisation with the out-of-domain corpus of some of the hyperparameters of the shallow RNN can be a reason for its good results, comparable with Transformer regarding the systems trained only with out-of-domain corpora, and similar to deep RNN when adding the clinical domain resources.

	dev	test
Shallow RNN (out-of-domain)	10.69	10.67
Shallow RNN (+in-domain)	23.57	21.59
Deep RNN (out-of-domain)	7.23	5.91
Deep RNN (+in-domain)	23.01	20.74
Transformer (out-of-domain)	10.92	10.55
Transformer (+in-domain)	**26.67**	**24.44**

Table 5: BLEU values (Basque-to-Spanish) for different architectures using a shallow RNN for backtranslation.

However, if we compare the results of the different architectures trained with all the available re-

[5] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl (Accessed on April 11, 2019.)

sources, we see that Transformer outperforms both RNNs by around 3 BLEU points in each evaluation set. Thus, we can say that the Transformer architecture is the optimal for our task of translating clinical texts from Basque to Spanish.

5.2 Backtranslation systems

After determining which is the best general architecture for our task, we compare the results of different backtranslation systems. First, we evaluate the performance of the systems used to translate the available EHRs in Spanish into Basque, using as a reference the same datasets employed for evaluating the different architectures. Table 6 shows the results of the tested backtranslation systems.

	dev	test
RBMT$_{bt}$	8.56	7.03
SMT$_{bt}$	10.30	8.75
Shallow RNN$_{bt}$	10.75	10.44
Transformer$_{bt}$	**11.30**	**12.04**

Table 6: BLEU values for different backtranslation systems (Spanish-to-Basque).

We observe that the values obtained with NMT systems are similar to the ones obtained in the other direction with the system trained out-of-domain, which is logical since we only added the dictionaries for training the backtranslation systems. The results of SMT are also similar, with a slightly lower score in the test set. The results for RBMT are even lower, which can be because BLEU underestimates the results of RBMT systems in general.

Finally, we present in Table 7 the results in the clinical domain of the systems trained with the best performing architecture (Transformer) using all the training corpora, changing the method used for backtranslating the EHRs in Spanish.

	dev	test
RBMT	22.98	21.91
SMT	22.78	21.43
Shallow RNN	26.67	24.44
Transformer	**27.70**	**25.61**

Table 7: BLEU values (Basque-to-Spanish) for Transformer architecture using different backtranslation systems.

We notice that using Transformer for backtranslation obtains the best results, gaining more than 1 BLEU point comparing with the same Transformer architecture using a shallow RNN for backtranslation. The results for RBMT and SMT are lower, but comparing to the BLEU values for the backtranslation systems (Table 6), we observe that in this case the results using RBMT are slightly better than the ones with SMT. Apart from the aforementioned possible underestimation of RBMT systems when calculating BLEU, we think that this could be because the RBMT system can translate words that corpus based systems cannot translate, adding more variability to the source language corpus.

5.3 Ensemble of best models

After evaluating the performance of different architectures and backtranslation systems, we evaluate the performance of an ensemble of the 3 systems obtaining highest BLEU values in the development set, which in this case correspond to 3 different models of the Transformer architecture, using Transformer as backtranslation system, saved after different number of iterations. Specifically, the models evaluated for the ensemble are those saved after 90,000, 160,000 and 180,000 iterations, obtaining 27.56 BLEU points with the first two models, and 27.70 BLEU points with the last one. Table 8 shows the results of the ensemble system, which we name IxaMedNMT-Transformer. We observe gains of 0.33 BLEU points in the development set and 0.11 BLEU points in the test set, comparing to the results of the single model that obtained the highest BLEU value in the development set.

	dev	test
IxaMedNMT-Transformer	28.03	25.72

Table 8: BLEU values (Basque-to-Spanish) for an ensemble of the best performing systems.

5.4 Translation example and error analysis

Finally, Figure 1 shows an example of a translation performed by the ensemble system whose BLEU values were shown in Table 8, along with the original sentence in Basque and the manual translation into Spanish used as a reference.

We observe that the generated translation is almost equivalent to the human translation, with only slight differences in some of the words (presents/with, complete/wide, stenoses/obstructs, part/region, etc.), but without changing the overall meaning of the original sentence in Basque.

Original sentence in Basque

azaleko izter-arteria-k buxadura zabala du, baina iragazkor dago Hunter-en eremu-raino,
superficial femoral artery-ERG obstruction wide has, but permeable is Hunter-GEN region-ALL,
'the superficial femoral artery has a wide obstruction, but it is permeable up to the Hunter region,...

bertan buxatu eta 3. eremu popliteo-an berriz ere iragazkor bihurtzen da.
there obstruct and 3rd region popliteal-LOC again also permeable becoming is.
it is obstructed there and becomes permeable again in the 3rd popliteal region.'

Manual translation into Spanish

arteria femoral superficial con estenosis amplia pero permeable hasta la zona de Hunter
artery femoral superficial with stenosis wide but permeable up-to the region of Hunter
'superficial femoral artery with wide obstruction but permeable up to the Hunter region...

donde se estenosa, y en la zona 3 poplítea se vuelve otra vez permeable.
where it stenoses, and in the region 3 popliteal it becomes another.F time permeable.

where it stenoses and becomes permeable again in the popliteal region 3.'

Translation by the IxaMedNMT-Transformer system

la arteria femoral superficial presenta una oclusión completa que se encuentra permeable hasta el
the artery femoral superficial presents a occlusion complete which is found permeable up-to the
'the superficial femoral artery presents a complete occlusion which is permeable up to the...

área de Hunter, donde se obstruye y se vuelve permeable en la 3ª porción poplítea.
region of Hunter, where it obstructs and it becomes permeable in the 3rd portion popliteal.

Hunter region, where it is obstructed and becomes permeable in the 3rd popliteal portion.'

Figure 1: Translation example by the IxaMedNMT-Transformer system, along with the original sentence in Basque and the manual translation into Spanish.

In a fast overview of the whole of the sentences translated from the development set, we have observed that for some of the long sentences, the translation ended abruptly without translating a few of the last words. We have tried to scale down the beam-width from 10 (optimised for the shallow RNN, kept in other architectures for fair comparison) to the default value of 5 to reduce the probability of generating the end-of-sentence token sooner than necessary, but the BLEU values in the development set did not improve as expected. We plan to test diverse values of length-normalisation and coverage-penalty coefficients to try to overcome this problem.

This phenomenon occurred especially in sentences with a lot of punctuation marks, usually containing a list of symptoms, diseases or drugs. Regarding the translation of rare words, like in this case drug names, we have observed very few errors where part of the word was not translated correctly due to the BPE word segmentation. In the future, we intend to perform a thorough analysis of the different types of errors encountered in the generated translations, with the aim of developing possible solutions to them.

6 Conclusions and future work

We have showed that it is possible to translate clinical texts from Basque to Spanish without clinical domain bilingual corpora. We have leveraged previous work in translation of clinical terminology into Basque (Perez-de-Viñaspre, 2017), described a method for creating artificial sentences based on SNOMED CT relations, and made use of available EHRs in Spanish. Given the multilinguality and rich structure of SNOMED CT, similar dictionaries and artificial sentences might be generated for other language pairs for which bilingual clinical corpora are not available.

Furthermore, we have tested our method with different NMT architectures and using diverse systems for backtranslation, including rule-based and statistical systems. We obtained the best results using Transformer for both general architecture and backtranslation systems, achieving 28 BLEU points in the development set through checkpoint ensembling, and showing a translation example.

We leave as future work the human evaluation of the best performing system, with the possibility of improving the corpora used for training and evaluation.

Acknowledgements: This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) FPI grant number BES-2017-081045, and projects BigKnowledge (BBVA foundation grant 2018), DOMINO (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE) and PROSA-MED (TIN2016-77820-C3-1-R, MCIU/AEI/FEDER, UE). We thank Uxoa Iñurrieta for helping us with the glosses.

References

Agirre, Eneko, Inaki Alegria, Xabier Arregi, Xabier Artola, Arantza Díaz de Ilarraza, Montse Maritxalar, Kepa Sarasola, and Miriam Urkia. 1992. XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology. In *Proceedings of the third conference on Applied natural language processing*, 119–125.

Alegria, Iñaki, Xabier Artola, Kepa Sarasola, and Miriam Urkia. 1996. Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, 11(4):193–203.

Ataman, Duygu, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108:331–342.

Bapna, Ankur, and Orhan Firat. 2019. Non-Parametric Adaptation for Neural Machine Translation. *arXiv preprint arXiv:1903.00058*

Barone, Antonio Valerio Miceli, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. *arXiv preprint arXiv:1707.07631*

Burlot, Franck, and François Yvon. 2019. Using Monolingual Data in Neural Machine Translation: a Systematic Study. *arXiv preprint arXiv:1903.11437*

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*

Currey, Anna, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, 148–156.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*

Etchegoyhen, Thierry, Andoni Azpeitia, and Naiara Pérez. 2016. Exploiting a Large Strongly Comparable Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Porotoroz, Slovenia.

Etchegoyhen, Thierry, Eva Martínez, Andoni Azpeitia, Gorka Labaka, Iñaki Alegria, Itziar Cortes, Amaia Jauregi, Igor Ellakuria, Maite Martin and Eusebi Calonge. 2018. Neural Machine Translation of Basque. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain, 139–148.

Heafield, Kenneth. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 187–197.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

International Health Terminology Standards Development Organisation IHTSDO. 2014. *SNOMED CT Starter Guide*. Technical report, International Health Terminology Standards Development Organisation

Jauregi, Inigo, Lierni Garmendia, Ehsan Zare, and Massimo Piccardi. 2018. English-Basque statistical and neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 880–885.

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

Kalchbrenner, Nal, and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709.

Kingma, Diederik P., and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 48–54.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In

Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, 177–180.

Koehn, Philipp, and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*

Labaka, Gorka. 2010. *EUSMT: incorporating linguistic information into SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation.*. PhD thesis, University of the Basque Country, Donostia, Euskal Herria.

Lee, Jason, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017*

Liu, Weisong, and Shu Cai. 2015. Translating electronic health record notes from English to Spanish: A preliminary study. *Proceedings of BioNLP 15*, 139–148.

Mayor, Aingeru. 2007. *Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz.*. PhD thesis, University of the Basque Country, Donostia, Euskal Herria.

Otegi, Arantxa, Nerea Ezeiza, Iakes Goenaga, and Gorka Labaka. 2016. A Modular Chain of NLP Tools for Basque, 93–100.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318.

Perez-de-Viñaspre, Olatz. 2017. *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.

Poncelas, Alberto, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *arXiv preprint arXiv:1804.06189*

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*

Sennrich, Rico, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357*

Stroppa, Nicolas, Decan Groves, Andy Way, and Kepa Sarasola. 2006. Example-based machine translation of the basque language. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, MA USA, 232–241.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Tang, Gongbo, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. *arXiv preprint arXiv:1808.08946*

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Zeiler, Matthew D. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*