Universidad del País Vasco Euskal Herriko Unibertsitatea

Basque-to-Spanish and Spanish-to-Basque Machine Translation for the health domain

Author: Xabier Soto García

Advisors: Gorka Labaka and Olatz Perez de Viñaspre

Co-advisor: Maite Oronoz

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

June 2018

Department: Computer Languages and Systems

Laburpena

Master Amaierako Lan honek medikuntza domeinuko euskara eta gaztelera arteko itzulpen automatiko sistema bat garatzeko helburuarekin emandako lehenengo urratsak aurkezten ditu. Corpus elebidun nahikoaren faltan, hainbat esperimentu burutu dira Itzulpen Automatiko Neuronalean erabiltzen diren parametroak domeinuz kanpoko corpusean aztertzeko; medikuntza domeinuan izandako jokaera ebaluatzeko ordea, eskuz itzulitako corpusa erabili da medikuntza domeinuko corpusen presentzia handituz entrenatutako sistema desberdinak probatzeko. Lortutako emaitzek deskribatutako helbururako bidean lehenengo aurrerapausoa suposatzen dute.

Abstract

This project presents the initial steps towards the objective of developing a Machine Translation system for the health domain between Basque and Spanish. In the absence of a big enough bilingual corpus, several experiments have been carried out to test different Neural Machine Translation parameters on an out-of-domain corpus; while performance on the health domain has been evaluated with a manually translated corpus in different systems trained with increasing presence of health domain corpora. The results obtained represent a first step forward to the described objective.

Contents

List of Figures

List of Tables

1 Project definition

The objective of this Final Thesis for the Language Analysis and Processing Master (HAP/LAP) is to analyze different techniques for Basque-to-Spanish and Spanish-to-Basque Machine Translation (MT) on the health domain. Specifically, distinct configurations of Neural Machine Translation (NMT) systems will be tested trying to adapt to the resources available for the health domain in Basque and Spanish languages.

Basque is a minoritised language, which has its reflection also in the Basque public health service, where mostly all of the health records are registered in Spanish so as to any doctor can understand them. Nowadays, if any patient wants to consult his or her health record in Basque, the translation of the health record is done by human translators on demand. With a view to guaranteeing the linguistic rights of all doctors and patients, the purpose of this project is to study different MT systems and parameters so as Basque speaking doctors are able to write in Basque without worrying about who can not understand the health records; and from the patients' point of view, to have access to their health records in the language they choose without waiting for a manual translation.

The increasing availability of health records in a digital format, commonly known as Electronic Health Records (EHR), makes possible this kind of MT techniques. However, the main handicap of this project is the lack of bilingual corpora for the health domain in Basque and Spanish. To alleviate this problem, different approaches will be tried such as i) inserting a medical dictionary to an out-of-domain bilingual corpus, ii) creating artificial sentences from a health-related ontology or iii) adding a health domain monolingual corpus along with its machine translation.

Taking into account that data from the health domain are extremely sensitive, all the personal information from patients have to be deleted from EHRs before using them for developing MT systems. In addition to that, comparing with other purposes of MT systems, in MT for the health domain accuracy of the results has to be even more important when choosing different techniques or parameters, so other aspects like computation time or hardware requirements will be less considered.

The thesis report has been organised as follows. In Chapter 2 we will describe the basic approaches to Machine Translation and introduce the fundamental parameters of Neural Networks that could be tested. In Chapters 3 and 4 we will explain the followed methodology and our approach to this problem. In Chapter 5 we will show the results obtained in each of the experiments, and finally, Chapter 6 will present some conclusions and suggest future developments in this area.

2 Antecedents

This chapter will be divided into two parts, in the first part (Section 2.1) we will briefly describe the different approaches for MT, mentioning the works for MT between Basque and Spanish for each of the approaches, and we will explain the basics of Neural Networks, paying special attention to the different parameters we will test. Furthermore, we will refer to previous work on automatic creation of Basque terminology for the health domain, which could be useful also for this project. Finally, we will describe different methods of evaluation for Machine Translation systems.

In the second part of this chapter (Section 2.2) we will specify the different approaches to overcome the main challenges for the objective of this project, namely the difficulty of performing NMT between linguistically different languages, as is the case of Basque and Spanish; domain adaptation for NMT, with the specificities of the health domain; and the aforementioned handicap of performing the NMT task with low resources

2.1 Machine Translation

Language is the most important way of communication between humans, and probably the main characteristic that distinguises us from other species. However, the existence of different languages is at the same time a communication barrier and a heritage to be preserved. In this context, the use of technologies plays a fundamental role to overcome this barriers and help maintaining language diversity.

Machine Translation is defined as the process to automatically translate a text from one natural language to another. As in other areas of computational linguistics, there are two main approaches for MT, one based on linguistic knowledge, usually referred as Rule-Based Machine Translation (RBMT); and another set of methods based on extracting information from already translated texts, grouped in the area of Corpus-Based Machine Translation (CBMT). From the latter, nowadays there are two dominant approaches, which are Statistical Machine Translation (SMT) and the aforementioned Neural Machine Translation (NMT). There exists also a third approach for CBMT called Example Based Machine Translation (EBMT), based simply on searching for patterns of the input sentence on a given bilingual corpus and recombine the corresponding translations to form an output sentence. In this chapter we will not describe the basics of EBMT, but some of the suggestions for future research make use of similar ideas to improve the results of other MT approaches. For a clearer explanation, in the following we will not refer to the two groups of systems (RBMT and CBMT) but to the specific techniques that will be considered in this work (RBMT, SMT and NMT).

2.1.1 Rule-Based Machine Translation

Rule-Based Machine Translation systems make use of the linguistic knowledge previously structured in the form of bilingual dictionaries, morphologic, syntactic and semantic analyzers, and a set of rules that define the relation between source and target languages.

These systems obtain better results for similar languages, and are difficult to maintain due to the expected capability of the set of rules to define all the possible sentence structures. Having said that, RBMT systems are predictable and easier to debug since they act as deterministic systems.

The translation process carried out by a RBMT system can be divided into 3 steps: analysis of the source language, tranfer of linguistic knowledge extracted from the input sentence, and generation of the output sentence following the linguistic rules of the target language. Depending on how deep is the analysis made, 3 kinds of RBMT systems can be defined: direct systems, based on a word-to-word translation from source to target language, transfer-based systems, in which some kind of linguistic analysis is carried out, and interlingual systems, which make use of a language-independent abstract representation of the input sentence. These ideas are plotted in the Vauquois triangle shown in Figure 1.

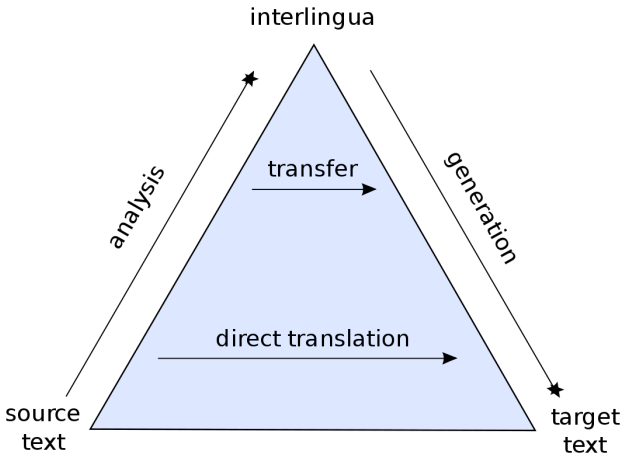

Figure 1: Vauquois triangle[1]

The first reference for Machine Translation between Basque and Spanish is Matxin (Mayor, 2007), a rule-based open source tool developed by the IXA group in the University of the Basque Country (UPV/EHU). An adaptation of Matxin to the health domain called MatxinMed also exists (Perez-de-Viñaspre, 2017), but will not be used for this work since it is implemented for English-to-Basque translation direction. Nowadays, we lack of any open sourced RBMT system using Basque as source language, so we can not work with it for domain adaptation.

2.1.2 Statistical Machine Translation

The increasing availability of more and more parallel corpora made possible for machines to automatically learn how to translate, based on different appearances of the same word

[1]Source: www.wikipedia.org

in different parts of a given bilingual corpus, and looking for the most probable translation around similar positions on the sentence in the target language.

SMT systems are composed of 3 distinct elements: 1) the Translation Model, which makes use of the parallel corpus to infer the relations between words in source and target languages; 2) the Language Model, which measures the probability of a given sequence of words in a monolingual corpus in the target language; and 3) the decoder, which performs the translation by means of some algorithms that make use of the statistical information extracted from Translation and Language Models. Figure 2 represents this kind of system.

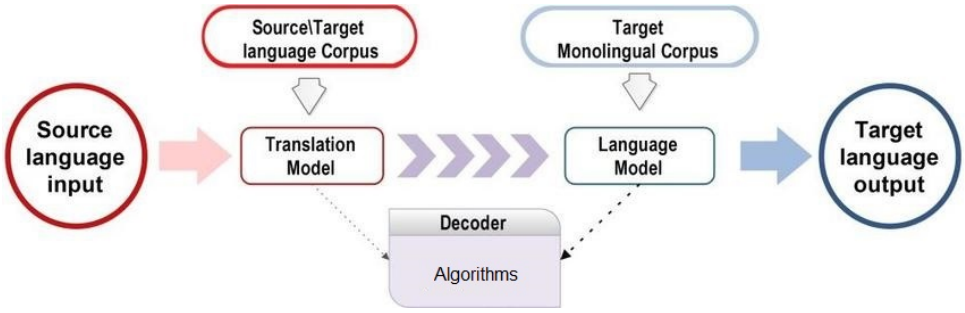

Figure 2: Components of a Statistical Machine Translation system[2]

When new input text is introduced into the system, translation probabilities of different possible output sentences are calculated and the ones with highest probability are given as output of the system.

Original SMT models used words as basic units for translation (Brown et al., 1993), but as the technique evolved more sophisticated systems based on phrases were developed (Koehn et al., 2003).

In general, SMT systems achieve better results when translating between languages with similar word ordering and morphology. On the other hand, ought to its dependence on statistics extracted from parallel corpora, SMT systems are unable to translate words that do not appear in the training corpus.

The reference tool for SMT between Basque and Spanish is EuSMT (Labaka, 2010). It is foreseen also to adapt EuSMT for the health domain (Perez-de-Viñaspre, 2017), but as for the techniques explored in this work, the unavailability of bilingual corpora act still as a handicap.

2.1.3 Neural Machine Translation

Neural Machine Translation is the result of applying the theory of Neural Networks to Machine Translation. The first works that suggested this possibility were Forcada and Ñeco (1997) and Castaño and Casacuberta (1997), but the limitation on computational capabilities did not allow to develop this area in that moment. It was not until more than

[2]Source: adapted from Elsherif and Soomro (2017)

15 years after when works by Kalchbrenner and Blunsom (2013) and Sutskever et al. (2014) recovered this idea with a real possibility of applying it.

The idea behind Neural Networks is to estimate complex functions simulating how the neurons in the brain work, where the signal each neuron emits to other neurons depends on the signals received by neighbouring neurons and some weights asociated to each of the connections with different neurons. Figure 3 shows a representation of these real neural networks (left), together with a simple artificial neural network (right), based on one layer of hidden units with respective weights (w) for each of the connections with units from input and output layers. Non-linear functions (δ) are used to outperform classical Machine Learning approaches and be able to, in theory, approximate any given function.

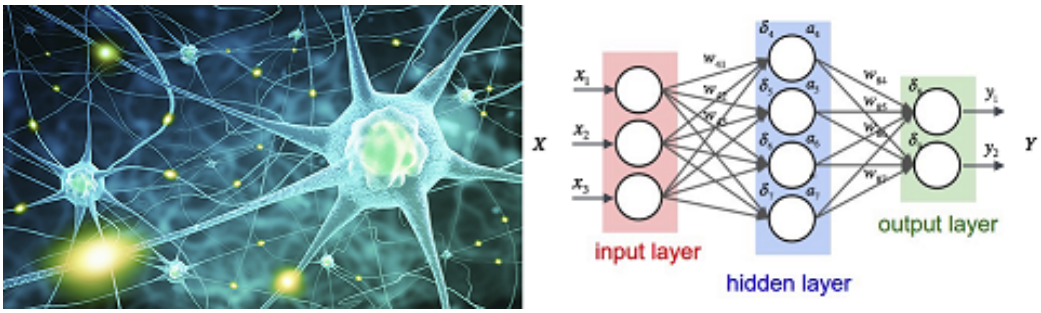

Figure 3: Example of real and artificial neural networks[3]

Usual configurations of Neural Networks for Machine Translation consist on what is known as encoder-decoder configurations, with one neural network such as the one shown in the right side of Figure 3 (with the possibility of having more than one hidden layer) for the encoder and another one for the decoder. The process of training a Neural Network consists then on 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 back-propagation (Rumelhart et al., 1986) until some loss function is minimised. Then the trained model can be tested with new input data.

The main characteristic of NMT compared to previous techniques is that NMT systems act as a black box that learn how to translate without making use of any linguistic or statistical information, just trying to mimic the abstract process of translation. For doing this, input text in source language is encoded into numerical values, representing word and sentence meanings as vectors, which then will be decoded into output sentences in target language.

Recently, NMT has shown to be the most effective system for Machine Translation (Cho et al., 2014), making some significant improvements like the inclusion of an attention-mechanism to automatically predict which are the most relevant words on a source sentence to be translated into the next output word (Bahdanau et al., 2014), or using word segmentation to improve the translation of rare words (Sennrich et al., 2015b).

[3]Source (left): http://www.neuraldump.net/2016/03/introduction-to-neural-networks/ Source (right): https://medium.com/@curiousily/tensorflow-for-hackers-part-iv-neural-network-from-scratch-1a4f504dfa8

These recent advances on NMT have also been tested for Basque-to-Spanish and Spanish-to-Basque Machine Translation (Etchegoyhen et al., 2018), already improving the results of SMT systems.

In the following points different characteristics of NMT systems will be defined, being some of them the parameters that will be tested in next chapters to try to improve the previous results in this area.

1) Architectures

The most general distinction we can make among different NMT systems is their architecture, that is, the way layers of neurons are arranged to encode or decode a given data. Convolutional Neural Networks (CNN) are characterised by sharing some weights among consecutive input data, so are better suited to process continuous data such as images. Meanwhile, Recurrent Neural Networks (RNN) are better adapted to sequences of variable length like text, since the information saved in one hidden unit is also processed by the next hidden unit together with the current input data. Figure 4 shows an example of a Recurrent Neural Network.

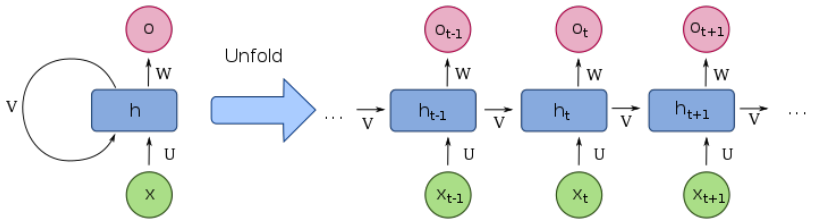

Figure 4: Example of a Recurrent Neural Network[4]

Recently, more complex architechtures like the Transformer (Vaswani et al., 2017) have been defined to achieve state-of-the-art results for some language pairs. In our approach, as in the majority of NMT systems, a RNN architecture will be used.

2) Attention-mechanism

Regardless of the neural network architecture, an attention-mechanism can be used to improve the performance of the system when the task requires it. For instance, in the case of CNNs, attention-mechanism can be used in image captioning systems to allow the decoder to focus on specific parts of the input image when generating a new output word.

Similarly, RNNs used for Machine Translation can be highly benefited from an attention-mechanism to focus on specific words from the input sentence when generating a new output word. This idea is similar to the allignment needed in SMT systems, as can be shown when representing the atention-weights for the different input and output word relations. Figure 5 shows an example extracted from Bahdanau et al. (2014) in which we can observe that each output word corresponds mainly to another input word.

[4]Source: www.wikimedia.org

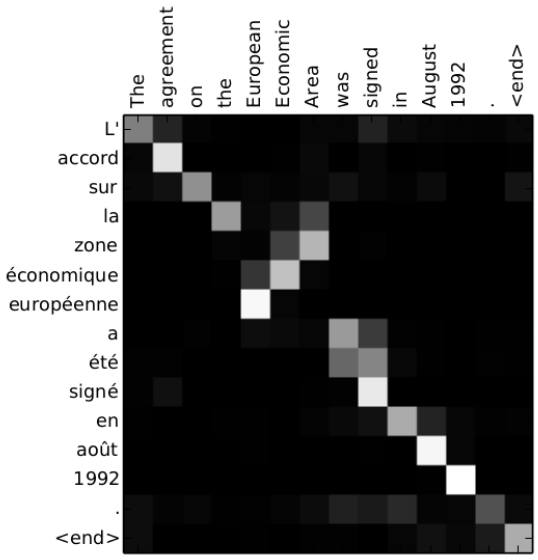

Figure 5: Attention-weights on a RNN used for MT (Bahdanau et al., 2014)[5]

3) Number of layers

Once a specific architechture is chosen, one direct way of augmenting the network complexity to try to improve the results is to increase the number of layers of the system. However, this is not always possible since it is limited by the computational capabilities.

In practice, most of the NMT systems use only 1 layer (with a specificity that will be described in the following section), and only researchers with access to much more powerful systems can test a higher number of layers. Two examples of this are Britz et al. (2017), in which they test different parameters of a NMT system including number of layers from 1 to 4, and the most advanced system used by Google for some language pairs which makes use of 8 layers (Wu et al., 2016).

In this work we will use a RNN with 1 layer.

4) Directionality

One limitation of RNNs is that in the moment of predicting a given word they can only access to the encoding of the sentence containing the meaning of previous words in the original sentence, but as we know, word ordering can differ between different languages, so it is necessary for the system to be able to read also future words from the input sentence. To overcome that, bidirectional RNNs are used, in which one layer reads the sentence in one direction and the other reads it in the opposite direction. This is the type of RNN that we will use in this project, so when we say that we use a RNN with 1 layer we mean one layer for reading from left to right and another one for reading from right to left. Figure 6 shows an example of a bidirectional RNN as the one used in this project.

[5]Source: https://blog.heuritech.com/2016/01/20/attention-mechanism/

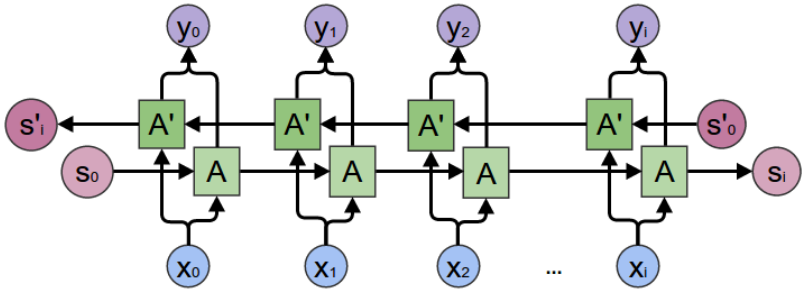

Figure 6: Example of a bidirectional RNN[6]

5) Unit type

Another drawback of using RNNs for Machine Translation is that, since the encoders accumulate the meaning of a sentence after reading every input word, it is difficult to maintain the long-distance dependencies between words that share some syntactic or semantic relation in long sentences. To overcome that, neurons can be replaced by more complex units which are able to remember/forget specific information from a given sentence.

The original approach to solve this problem is based on Long Short-Term Memory (LSTM) units (Hochreiter and Schmidhuber, 1997), but there exist also some simpler adaptations of them, standing out from all of them the one called Gated Recurrent Units (GRU) (Cho et al., 2014). In this project we will test both unit-types LSTM and GRU, whose schemes are shown in Figure 7.

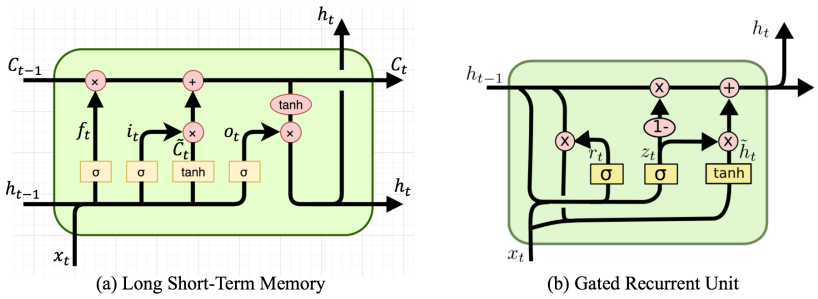

Figure 7: Schemes of LSTM and GRU unit-types[7]

6) Number of units

For each layer of the Neural Network, a number of units per layer will be specified depending on the computational capabilities. In this project, a fixed number of 1024 units per layer will be used.

[6]Source: http://colah.github.io/posts/2015-09-NN-Types-FP/
[7]Source: https://isaacchanghau.github.io/post/lstm-gru-formula/

7) Basic elements

Given a specific NMT architecture, the most general distinction we can make is which are the basic elements to be used as input for each unit of the neural network. The most logical and simple approach could be to use just words, but recent studies have shown that dividing the words into subwords can improve the translations of unknown words (Sennrich et al., 2015b), specially for highly inflected languages as Basque.

A more complex technique called Character-based Machine Translation (Costa-Jussà and Fonollosa, 2016) has been taken into consideration because it could be appropiate for morphologically rich languages as Basque, but has been discarded because of not achieving good enough preliminary results (Etchegoyhen et al., 2018).

In this work we will use the aforementioned word segmentation method (Sennrich et al., 2015b), commonly known as BPE word segmentation.

8) Vocabulary-size

As mentioned before, each element of input and output text is represented with numerical values, commonly known in Natural Language Processing (NLP) as embeddings. During the preprocessing, a number of embeddings will be created to represent the whole corpus of each source and target languages, constrained by an input parameter that determines the maximum number of embeddings that could be created.

When applying this preprocessing, it is possible to share the vocabulary for both languages, in which case the maximum vocabulary-size will limit the number of embeddings for one of the languages, using some of them to represent tokens for the other language. This parameter will be conditioned by the computational capabilities, and for this project a maximum vocabulary-size of 90,000 (sub)words will be used, with the option of sharing the vocabulary between languages activated.

9) Embedding-size

For representing each of the tokens from the input and output texts, vectors of a specified length will be used. In theory, embeddings of higher size will represent better the semantics of each word, and can also obtain better results for NMT (Britz et al., 2017). However, the embedding-size will also be conditioned by the computational capabilities, requiring more memory to train systems with higher embedding-size. For this project, embedding-sizes of 500, 512 and 1024 will be tested.

10) Optimization

The term optimization refers to the process carried out to search the minimum error during the training process. Different techniques are used for this, with the possibility or not of reaching the global minima in a shorter or longer time. Gradient descent, based on the derivatives of the functions implemented by the Neural Network, is the simplest method

used for optimization, but more complex methods using other parameters as momentum and learning-rate have demonstrated to be more efficient.

In this project, two of these methods will be tested, named Adadelta (Zeiler, 2012) and Adam (Kingma and Ba, 2014).

11) Batch-size

When it comes to updating the neural network parameters, an important feature is how many training samples are taken into account between each update. In the simplest scheme, with Stochastic Gradient Descent, all the neural network parameters are updated after calculating the error for each training sample, which makes the training process unnecessarily slow. On the other hand, updating the parameters after calculating the errors for all the training dataset would require great amounts of memory, so in practice the training set is divided into mini-batches of specific size and the parameters are updated after training the network with each mini-batch.

In this work, batch-sizes of 30, 32 and 64 will be tested.

12) Learning-rate

The final step to update a given parameter is to substract its derivative to the current value of the parameter. However, doing this can result in overfitting the value of the parameters to the sample(s) considered in this step, so a parameter named learning-rate is defined to measure to which extent have every sample or group of samples to be taken into account when updating the parameters. With mostly used optimization methods like the ones tested in this project the learning-rate is modified during the training process to make it faster, so the specified value of learning-rate only refers to its initial value.

In this work, an initial learning-rate of 0.0001 is used.

13) Drop-out

As mentioned when defining learning-rate, one problem of neural networks is that their parameters can be too much adapted to the data used in training process, producing what is known as overfitting. Different regularization techniques have been developed to overcome this problem, being drop-out the most common one. This method is simply based on randomly dropping units from the neural network during training process to avoid they became too dependent on the training data (Srivastava et al., 2014), and as other regularization techniques, can improve the results when overfitting is detected.

In this work, drop-out is not used.

14) Beam-width

All the previously described parameters are related to the training process, but there are also some parameters associated with the translation process. At each step of the decoding process, the softmax layer will output the probability for each word of being the

output word in this position, but it does not always happen that the most probable word in a specific position is the one that will correspond to the most probable translated sentence. In order to expand the search of the most likely output sentence, a parameter called beam-width is defined referring to the number of possible outputs that will be considered when choosing each of the most probable outputs of the decoder.

In this work, beam-widths of 6 and 10 will be tested.

15) Length-normalization and coverage-penalty

To calculate the probability of a given output sentence, the probability of each output word will be multiplied, so the decoder will tend to choose sentences of shorter length. This will make the results of translating longer sentences worse comparing to the results of translating shorter sentences, so a parameter called length-normalization is included to compensate this. In addition, another parameter called coverage-penalty is used to favour sentences that cover all the words from the input sentence, with some improvements made to the attention-module (Tu et al., 2016).

In this work, length-normalization is used with a value of 1, and coverage-penalty is not used.

2.1.4 Automatic creation of Basque terminology for the health domain

Apart from explaining the basics of different Machine Translation systems, it is important to mention the previous work done specifically in Basque language for the health domain. As said before, there is a lack of Basque corpora in this domain, thus a first step to someday have the possibility of having this corpora is to start by creating Basque terminology for the health domain, so doctors and different health workers have a reference when writing clinical texts.

Around the globe, there are different terminological databases and ontologies that health workers use as reference, and from all of them SNOMED CT was chosen in Perez-de-Viñaspre (2017) to automatically translate its terms to Basque. For doing that, different resources and techniques were used, and here we will only mention the ones that could be helpful for this project.

Regarding the resources used, information from different dictionaries was stored in a database named ItzulDB that was later used to carry out the translation process, compiling information from different sources such as Elhuyar Science and Technology dictionary, UPV/EHU human anatomy atlas and nursery dictionary, International Classification of Diseases dictionary and a health administration related dictionary.

With regard to the techniques used in Perez-de-Viñaspre (2017) that could be applyed to this work, it is necessary to highlight the already mentioned RBMT system Matxin-Med. This system, used in the final step of Perez-de-Viñaspre (2017) to automatically create Basque health terminology from SNOMED CT, is designed for Basque and English language pair, so it will not be used in this work but will be taken into account in the future for being adapted to Basque/Spanish Machine Translation.

Finally, one of the possible future works listed in Perez-de-Viñaspre (2017) is to obtain a stable version of SNOMED CT in Basque, with its content being reviewed by medical experts to certify its validity. Indirectly, this would be highly valuable for the objective of this project as long as it will be very helpful to start creating the bilingual corpus needed for Statistical and Neural Basque/Spanish Machine Translation for the health domain.

2.1.5 Evaluation in Machine Translation

Comparing to evaluation in other NLP tasks, evaluation in Machine Translation has to deal with the fact of having more than one possible correct translation. To overcome this, more than one reference translation can be used, with the consequent complexity added to the system. However, no matter how many correct translations are included in the set of reference translations, there could be always another possible translation, so evaluation in Machine Translation always carries some degree of uncertainty.

There are two main forms of evaluating Machine Translation systems: human and automatic. Usually, human evaluation is based on asking some people (who can be experts or not) to score some characteristics of the translation within a given range of possible values. The most common evaluated aspects are fluency, representing the naturalness of a given output sentence; and adequacy, representing how much of the information contained in the reference translation(s) is included in the output sentence. Another simpler approach for human evaluation consists just in choosing the best translation among a given set of translations, usually obtained with different systems. Human evaluation methods have the advantage of giving linguistic information of the errors made by the system, but they have the drawback of being inherently subjective and specially costly.

On the other hand, automatic evaluation consists in using some algorithm to automatically compare a given output sentence with an available reference translation. These methods have the advantage of being cheaper, faster and providing reproducible results, but not always being representative of the translation quality. The most used method for automatic evaluation of MT systems is BLEU (Papineni et al., 2002), which basically consists in counting the number of consecutive words that appear in both reference and system translation, with some corrective measure for too short outputs. In this project, we will use BLEU to compare different NMT systems and finally we will design a human evaluation method that would be carried out in the future on the best systems according to automatic evaluation.

2.2 Related work

When approaching the objective of this work, that is, building a NMT system between Basque and Spanish languages for the health domain, there are several perspectives that have to be taken into account, which they can be mainly divided into three areas: NMT between linguistically different languages, as is the case of Basque and Spanish; domain adaptation for NMT, with the specificities of the health domain; and the aforementioned handicap of performing the NMT task with low resources. In this Section, we will mention

separately the diverse related works in each of these areas, even if in some cases the developed techniques respond at the same time to more than one of the characteristics of the described problem.

2.2.1 NMT between linguistically different languages

Despite most Basque language speakers live in a territory surrounded by Spanish or French speakers, and almost all of the Basque speakers use also Spanish or French language in their everyday lives, the characteristics of Basque language are very different from their neighbouring languages. Indeed, Basque is an isolated language in the sense that there is no other language that can be linguistically related to it apart from the terms imported from other languages. Thus, while Spanish is a latin derived language sharing some characteristics with other European languages, Basque has its own characteristics that need specific treatment when approaching the NMT task.

In few words, 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 specially complex, including morphemes that add information about the subject, object, number, tense, aspect, etc. Furthermore, the order of the sentences is relatively free, which makes the development of NMT systems for Basque a specially challenging task, particularly for evaluation purposes.

In a very recent work, Etchegoyhen et al. (2018) show that better results can be obtained with NMT for Basque than with the traditional RBMT or SMT techniques. Specifically, they approach the problem derived by a complex morphology testing different word segmentation methods, from linguistically motivated ones to the well known BPE word segmentation method (Sennrich et al., 2015b). They also tested the character-based NMT (Lee et al., 2016), but in this case the results were worse than expected for a highly agglutinative language as Basque. For the sentence order variability, on the one hand, they manually created a second reference for the test set that accounted for this word order variability; and on the other hand, they tested different models with different values for length-normalization and coverage-penalty, based on the previous work by (Wu et al., 2016).

In another recent work, Passban et al. (2018) approach the complexity of dealing with morphologically rich languages as target language on NMT task by combining a traditional RNN encoder with BPE word segmentation on the source side (in this case, for English language) and a character-based decoder supplemented with morphology tables that are used in a similar way of attention modules to choose the most likely next output character depending on the morphologic information of the target language (in this work, German, Russian and Turkish).

2.2.2 Domain adaptation for NMT

In the case of NMT, despite the overall results are nowadays better than the ones obtained with SMT (Bojar et al., 2016), when the output of both systems is evaluated it is observed

that NMT systems generate sentences with better fluency, thus sounding more natural, while SMT systems are still better in terms of precision measured above each of the generated words (Koehn and Knowles, 2017). Since the NMT approach uses word embeddings to represent input and output words, this worse precision is usually not a big problem provided that the generated words are similar to or related with the correct output word, but in the specific case of NMT for the health domain, in which precision is probably the most important aspect to preserve, some action must be taken to improve this aspect.

In this sense, the recent works by Gu et al. (2017) and Zhang et al. (2018) represent a promising research area to the specific task of domain adaptation for NMT, specially in cases in which the sentences that form the training corpus tend to be similar, as is the case of health records. Both cited works develop the same basic idea, look for similar sentences to the input sentence before translating it, but differ in the way this sentence similarity information is used. While Gu et al. (2017) use this information to add the k most similar sentences to the training corpus, Zhang et al. (2018) simplify this process just using the sentence similarity scores to rescore the possible output sentences before choosing the desired output. Both works are tested with legal domain corpora, which is characterised by having similar sentences as happens in health domain documents, obtaining significant improvements up to 6 BLEU points.

2.2.3 NMT with low resources

Finally, we will refer to the specific task of NMT when low resources are available, as is the case with minoritised languages as Basque, and more dramatically when the domain is constrained like in this project oriented to the health domain. Even if NMT started to obtain competitive results as large amounts of bilingual corpora in digital format became available, this is still not the case for languages used by less people like Basque, despite the enormous efforts of the Basque speaking community to generate this digital content, compile it, and make it available for everyone who needs to make use of it. To overcome this problem, shared with the majority of languages that are trying to survive in a highly connected world with a few languages dominating the majority of digital content, a very interesting research area is starting to be developed trying to perform the NMT task when no bilingual data is available.

In this respect, the works by Artetxe et al. (2017) and Lample et al. (2017) have proved to obtain good results in the new born area of Unsupervised Machine Translation, that is, with no use of any bilingual corpora. These works, published only with one day difference, make use of the information intrinsic to a given language by exploiting the information contained on the word embeddings created with the available monolingual corpora, and then studying the best ways to relate the embedding maps created for each of the languages to be used in the translation process. Both works suppose a milestone that changes the traditional paradigm that bilingual corpora is needed to perform the NMT task, but as expected, they still not obtain state-of-the-art results when comparing to NMT systems that make use of bilingual corpora.

In consequence, nowadays there are other well established techniques that help to

achieve competitive results when low resources are available, as is the case of transfer learning (Zoph et al., 2016), based on first training a system with a big enough general corpus and then fine-tune some of the parameters training the system again with a smaller corpus that can be from a specific domain; or backtranslation (Sennrich et al., 2015a), based on including a monolingual corpus and its automatic translation to a similarly in size bilingual training corpus. 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), as is our case for Basque/Spanish Machine Translation for the health domain.

3 Methodology

In this section, we will present the system and resources that will be used to carry out the experiments, along with a brief description of the equipment that will be used for the NMT training and evaluation.

3.1 System

Nowadays, most of the NMT researchers publish their code so anyone can reproduce their experiments, apply it to different language pairs, or even suggest improvements to the developed code. In this project, we will use the system known as Nematus (Bahdanau et al., 2014), which implements the aforementioned attention-mechanism and makes use of Theano library, based on Python. Compared to newer systems like Sockeye (Hieber et al., 2017) that can implement or even combine different architectures (CNN, RNN and Transformer), Nematus has the disadvantadge of being restricted to RNNs, but we have chosen it for being one of the systems that achieves state-of-the-art results on NMT (Sennrich et al., 2016), and having been previously used for some of the projects developed in IXA research group.

3.2 Resources

3.2.1 Corpora

As stated in the introduction, there is a lack of health domain bilingual corpora for Basque and Spanish languages, so for most of the experiments an out-of-domain and big enough corpus will be used.

Specifically, the basic corpus that will be used to train the NMT systems is formed by a total of 4.5M sentences in TrueCase format. 2.4M of them are a 3 times repetition of a news domain corpus, while the rest 2.2M sentences are from an out-of-domain corpus. Without counting the repeated corpus, the effective data expressed in tokens would be 102M tokens in Spanish and 72M tokens in Basque. 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).

Furthermore, a Spanish monolingual corpus from the health domain will be included to the previous corpus along with its automatic translation, following the work by Sennrich et al. (2015a). This corpus is made up of 2 corpora containing real health records from the hospitals of Galdakao-Usansolo and Basurto. The first one consists of 142,154 files compiled during 5 years (2008-2012) with a total of 52M tokens, while the second one is composed of 189,623 files collected in 2014 with a total of 57M tokens. Table 1 shows the first 10 sentences of the Spanish monolingual corpus that will later be translated.

First sentences of the Spanish monolingual corpus from the health domain
EKG: rítmico a 65 x
también refiere que desde la operación ha incrementado la dosis de noctamid hasta 2 mg al día y desde hace unos 3 meses bajó la dosis a 0-0-1.5 mg por la noche
- Deterioro cognitivo leve en seguimiento por Neurología
no bebedor ni otros hábitos tóxicos
BRDHH e intervalo PR normal
abdomen: a su ingreso en planta normal, ruidos normales
tacto RECTAL: heces blandas en ampolla rectal
evolucion: durante su ingreso además de la historia de alcoholismo ya reflejada, se evidencia un trastorno del estado de animo de carácter crónico, con sentimientos de vacío, impulsividad y alteración en relaciones interpersonales
no se aprecia flujo anterógrado
refieren que ha presentado múltiples episodios de infecciones urinarias tratadas sin problemas con augmentine

Table 1: First 10 sentences of the Spanish monolingual corpus from the health domain

3.2.2 Dictionaries and ontologies

Before including the health-related corpus, different experiments will be tried progressively adding health domain information extracted from dictionaries and ontologies to the previous out-of-domain corpus. For these experiments, the work developed in Perez-de-Viñaspre (2017) regarding SNOMED CT ontology will be taken as a reference.

For the first of the experiments, a dictionary will be built with all the created Basque terms and their respective Spanish counterparts stored in ItzulDB for translating into Basque the terms included in SNOMED CT. These Basque terms were automatically created, and is expected that their validity could be soon certified by health domain experts. For many of the Spanish terms refering to a specific SNOMED CT concept, more than one possible Basque term was created; so in total, the dictionary used for this experiment will have 151,111 entries corresponding to 83,360 unique Spanish terms. As a sample, Table 2 shows the first 10 elements of the dictionary.

For the second of the experiments, artificial sentences will be created making use of the relations specified on the SNOMED CT ontology. Specifically, the Snapshot release of the international version on RF2 format of the SNOMED CT delivery from 2017 July 31st will be used. For the sentences to be representative, the most frequent active relations will be taken into account, only considering the type of relations that appear more than 10,000 times. Table 3 shows the most frequent active relations along with their respective number of appearances.

Spanish term	Basque term
órgano copulador	organo kopulatzaile
dionisiaco	dionisiako
desfile	desfile
miasis ocular	begi-miiasia
candidiasis oral	ahoko kandidiasi
wolframio	wolfram
wolframio	W
recaudador	zergari
recaudador	jasotzaile
recaudador	biltzaile

Table 2: First 10 elements of the dictionary created using SNOMED CT Spanish terms and automatically created Basque translations

3.2.3 Health record models and manual translations

All of the aforementioned health-related resources will be used to enrich the training corpus, but for evaluating the performance of the system on the health domain, an additional corpus will be used. Specifically, a total of 42 health record models of diverse specializations written in Basque by doctors of Donostia Hospital (Joanes Etxeberri Saria V. Edizioa, 2014), and their respective manual translations into Spanish carried out by a bilingual doctor will be used as reference for evaluating the systems' performance on the health domain. These original health record models in Basque are written in a correct and well suited language, which makes them valuable not only for MT tasks but also as a model for Basque speaking doctors that want to start writing health records in Basque.

After aligning the sentences obtained from this EHRs and their respective manual translations, we will have a bilingual corpus consisting of a total of 2,076 sentences, that will be randomly ordered and equally divided into 1,038 sentences for the development (dev) set and another 1,038 sentences for the test set. As a sample, Table 4 shows the first 10 sentences that will be used for evaluation in Spanish and Basque.

Relation	**Number of appearances**
is a	502,459
Finding site	185,463
Associated morphology	159,155
Method	156,032
Procedure site - Direct	68,243
Procedure site	62,322
Part of	42,596
Interprets	42,205
Causative agent	21,132
Direct morphology	20,406
Procedure site - Indirect	20,066
Has active ingredient	19,294
Has interpretation	18,183
Temporal context	13,573
Subject relationship context	13,332
Occurrence	12,956
Direct substance	12,783
Pathological process	11,737
Has manufactured dose form	11,388

Table 3: Most frequent active relations and number of appearances on SNOMED CT

Basque sentence	**Spanish sentence**
tratamendua	tratamiento
abortuak : 1	abortos 1
lehenengo sintomatologia	primera sintomatología
fibrinolisiaren ondoren egoera klinikoa ez da askorik aldatu	la situación clínica después de la fibrinólisis no cambia sustancialmente
hipertentsioaren aurkako tratamenduarekin hasi da, tentsioak neurri egokian mantenduz ; hipergluzemiarako joera antzeman da egonaldian	al mismo tiempo tratamientopara normalizar la HTA, hiperglucemia y dislipemia
ebakuntza aurreko azterketa normala izan ostean, 2012-08-20an operazioa egin da	tras ser normal la exploración preoperatoria se opera el 20-08-2012, practicándose:
Dismetriarik ez	no dismetría
miaketa oftalmologikoa normala	examen oftalmológico normal
EKG: erritmo sinusala, 103 tau/min	EKG-ritmo sinusal 103/minuto
ez du botaka egin	no vómitos

Table 4: First 10 sentences that will be used for evaluation in Basque and Spanish

3.3 Equipment

NMT training requires big computational capabilities, with a lot of derivatives to be calculated and parameters to be updated at a time. To do all the calculations effectively, parameters are vectorised and Graphics Processing Units (GPU) are used because of their apropriateness to work with matrixes. In this project, two different GPU servers will be used: the first one, called Arina, consists mainly on a Tesla K40 GPU with 12 GB of RAM. This resource, external to IXA group, sends you a notification when a given job is finished, including the required computation time. The second server, exclusive for IXA group, is called Mamarro, from where a Titan Xp GPU with 12 GB is used. This system does not let you know the required computation time, but even if the specified memory capacity is the same, it is significantly faster than the previous one, resulting in around 2 days for training a model comparing to 7 to 9 days for the same job with Arina.

4 Our approach

In this section we will describe our approach for the experiments to be carried out in this project, which will be divided into two parts: the first part will consist on training different NMT systems on the bilingual out-of-domain corpus described in Section 3.2.1, changing one parameter at a time, and choosing the one which obtains best BLEU evaluation results for following experiments. In the second part, the health domain corpus defined in Section 3.2.3 will be used to automatically evaluate the best system according to the experiments carried out in the first part; and after this, health-related resources in the form of a dictionary, sentences created from SNOMED CT, or a Spanish monolingual corpus (described in Section 3.2) will be progressively added to measure their effect on the translation quality for the health domain. Finally, since automatic evaluation metrics do not always reflect perfectly the translation quality, a human evaluation method will be designed for the systems with best automatic evaluation numbers.

4.1 NMT parameters test

The corpus used for this part of the project will be the bilingual one specified in Section 3.2.1, with a total of 4.5M sentences. From this corpus, 2,000 sentences will be extracted for the dev set and another 2,000 for the test set; but after manually inspecting the correction of the sentences included in these sets, 1,994 sentences will be used as dev set and 1,678 sentences as test set. The rest of the sentences (4,530,683) will be used to train the system.

The starting point for this part of the project will be the NMT system developed for the Modela project (Etchegoyhen and Labaka, 2017), whose basic parameters, including the ones that will be tested in this work, are shown in Table 5.

When choosing the parameters to test, various sources were consulted, including several articles and online courses, but most of the parameters and their possible optimal values were taken from Britz et al. (2017). Table 6 shows all the parameters that will be tried and their respective values, in the same order in which they will be tried.

The rest of the parameters will remain as specified in Table 5, and all the experiments will be carried out for both translation directions Basque-to-Spanish and Spanish-to-Basque. After comparing the results for different values of each parameter, the one with higher BLEU values on the test dataset will be chosen for the next experiment, and only if the results are significantly different for each translation direction a different parameter value will be selected for each direction.

The experiments for optimization, unit-type, beam-width and batch-sizes of 30 and 32 will be carried out in Arina, so training time will also be included within the results. On the other hand, the experiments for embedding-size and a batch-size of 64 will be carried out in Mamarro, with no information about computation time.

Parameter	Value
Architecture	RNN
Number of layers	1
Directionality	Bidirectional
Unit type	GRU
Number of units	1024
Basic elements	Subwords (BPE)
Vocabulary-size	90,000 (shared)
Embedding-size	500
Optimization	Adadelta
Batch-size	30
Learning-rate	0.0001
Drop-out	Not used
Beam-width	6
Length-normalization	1
Coverage-penalty	0

Table 5: Parameters of the baseline system

Parameter	Values
Optimization	Adadelta / Adam
Unit type	GRU / LSTM
Beam-width	6 / 10
Batch-size	30 / 32 / 64
Embedding-size	500 / 512 / 1024

Table 6: Parameters tried in this project

4.2 Evaluation on the health domain

After testing different parameters on NMT systems trained with an out-of-domain corpus, we will replace the evaluation corpus for the one specified in Section 3.2.3, and subsequently add health domain corpora to the training corpus to evaluate their respective influence on health domain Machine Translation. These experiments will be developed for both Basque-to-Spanish and Spanish-to-Basque translation directions, except for the one including the Spanish monolingual corpus and its translation into Basque, that will be performed only for Basque-to-Spanish translation direction since the automatically translated corpus can not be taken as target training corpora, as it is stated in the original paper describing the backtranslation technique (Sennrich et al., 2015a). For these experiments the GPU server known as Mamarro will be used, so no information about computation time will be provided.

1) Using the out-of-domain corpus

For the first of these experiments, we will just choose the system that achieves best BLEU results on the out-of-domain corpus and evaluate it with the health domain corpus specified in Section 3.2.3.

2) Including a health-related dictionary

Then, we will start to add different health-related resources to measure their contribution to a better translation. For the first of these experiments, the information extracted from the dictionary specified in Section 3.2.2 will be used. For the results to be comparable with the previous that only uses an out-of-domain corpus, the preprocessing applied after including the dictionary will be the same, consisting of tokenization, TrueCase formatting and BPE word segmentation.

3) Including artificial sentences created from SNOMED CT

For the second of the health-related experiments, artificial sentences created from the relations on SNOMED CT ontology presented in Section 3.2.2 will be used. The reason for adding these sentences apart from specific terms 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 will first define two sentence models for each of the most frequent relations specified in Table 3, whose values are shown in Tables 7 and 8.

Taking these sentence models as a reference, and using the concepts from SNOMED CT as possible values of X and Y in Tables 7 and 8, for each of the concepts (X) concerning a unique pair of Basque and Spanish terms, we will randomly choose one of the relations that this concept has on SNOMED CT, restricting its possible values to the most frequent relations specified in Table 3 and omitting the relations with terms (Y) that are not among the translated ones. Finally, we will randomly choose one of the sentence models for this specific relation, taking a pair of sentences either from Table 7 or Table 8.

Once these artificial sentences are created, the models associated with each of the relation types will be reviewed, and in case the sentences are not appropiate for the majority of specific X or Y terms and it is not possible to define a sentence model that can be valid for all the possible X or Y values, the sentences associated with these relation types will be removed.

As a result of this process, we first have to state that when randomly choosing one relation for each of the terms available in Basque and Spanish, none of these relations corresponded to the 'Subject relationship context' type. Furthermore, as stated before, after reviewing the automatically created sentences, the ones corresponding to the relations 'Occurrence' and 'Direct substance' were removed for not being appropiate for the majority of specific terms used in each sentence, and not being possible to redefine a sentence model valid for the diverse terms that appeared for these relation types.

Finally, for applying the morphological inflections to the specific X or Y terms needed in some of the described sentences in Basque, a transductor will be applied following the

inflection rules defined in the transductors of Xuxen spelling corrector (Agirre et al., 1992). After this, a total number of 363,958 sentences were added to the corpus including the out-of-domain corpus and the health-related-dictionary, and the same preprocessing carried out when adding the dictionary will be applied before training the NMT system.

Relation	**Sentence model in Basque (I)**	**Sentence model in Spanish (I)**
is a	X Y da	X es Y
Finding site	X Yn gertatzen da	X ocurre en Y
Associated morphology	X Y moduan gertatzen den gaixotasuna da	X es una enfermedad que ocurre en forma de Y
Method	X Y behar duen prozedura da	X es un procedimiento que requiere de Y
Procedure site - Direct	X Y zuzenean ukitzen duen prozedura da	X es un procedimiento que afecta directamente a Y
Procedure site	X Y ukitzen duen prozedura da	X es un procedimiento que afecta a Y
Part of	X Yren parte bat da	X es una parte de Y
Interprets	X Y interpretatzen duen aurkikuntza da	X es un hallazgo que interpreta Y
Causative agent	X Yk eragindako gaixotasuna da	X es una enfermedad causada por Y
Direct morphology	X metodoa aplikatzean Y zuzenean ukitzen da	Al aplicar el método X se afecta directamente a Y
Procedure site - Indirect	X Y zeharka ukitzen duen prozedura da	X es un procedimiento que afecta indirectamente a Y
Has active ingredient	X produktuak Y substantzia dauka	El producto X tiene la sustancia Y
Has interpretation	X aurkikuntza Y bezala interpretatzen da	El hallazgo X se interpreta como Y
Temporal context	X Y kokatzen da	X se sitúa Y
Subject relationship context	X egoerak Y pertsonari eragiten dio	La situación X afecta a la persona Y
Occurrence	X lehenengoz Yn azaldu zen	X apareció por primera vez en Y
Direct substance	X prozedurak Y substantzia erabiltzen du	El procedimiento X utiliza la sustancia Y
Pathological process	X gaixotasunak Y suposatzen du	La enfermedad X supone Y
Has manufactured dose form	X produktua Y bezala banatzen da	El producto X se reparte como Y

Table 7: Sentence models used to create the artificial sentences (I)

Relation	Sentence model in Basque (II)	Sentence model in Spanish (II)
is a	X Y mota bat da	X es un tipo de Y
Finding site	X Yn aurkitzen da	X se encuentra en Y
Associated morphology	X Y forma hartzen duen gaixotasuna da	X es una enfermedad que toma la forma de Y
Method	X prozedurak Y suposatzen du	El procedimiento X supone Y
Procedure site - Direct	X prozedura praktikatzean Y zuzenean ukitzen da	Al practicar el procedimiento X se afecta directamente a Y
Procedure site	X prozedura praktikatzean Y ukitzen da	Al practicar el procedimiento X se afecta a Y
Part of	X Yren parte da	X forma parte de Y
Interprets	X aurkikuntzak Y interpretatzen du	El hallazgo X interpreta Y
Causative agent	X Yren eraginez sortutako gaixotasuna da	X es una enfermedad causada por efecto de Y
Direct morphology	X metodoak Y zuzenean ukitzen du	El método X afecta directamente a Y
Procedure site - Indirect	X prozedura praktikatzean Y zeharka ukitzen da	Al practicar el procedimiento X se afecta indirectamente a Y
Has active ingredient	X Y substantzia daukan produktua da	X es un producto que contiene la sustancia Y
Has interpretation	X aurkikuntzak Y interpretazioa du	El hallazgo X tiene la interpretación Y
Temporal context	X denboran Y kokatzen da	X se sitúa temporalmente Y
Subject relationship context	X egoera Y pertsonari dagokio	La situación X le correponde a la persona Y
Occurrence	X Yn agertu zen	X surgió en Y
Direct substance	Y substantzia X prozeduran erabiltzen da	La sustancia Y se utiliza en el procedimiento X
Pathological process	X Y moduan agertzen den gaixotasuna da	X es una enfermedad que aparece en forma de Y
Has manufactured dose form	X produktua Y moduan banatzen da	El producto X se reparte en forma de Y

Table 8: Sentence models used to create the artificial sentences (II)

4) Including a monolingual corpus and its translation

For this part of the project the EHRs included in the Spanish monolingual corpus specified in Section 3.2.1 will be used. These EHRs will be first preprocessed to have 1 sentence in each line and then the order of the sentences of the set of EHRs will be randomly changed to contribute to a better anonymization of the information included in each of them. For making the translation process faster, repeated sentences will be removed from the corpus before translate it, resulting in a total of 2,023,811 sentences that will be added to the previous corpus. For translating this sentences into Basque, the system including a health-related dictionary will be used.

Due to an error in the preprocessing, the training corpus used to perform the translation will be slightly different to the one described in Section 3.2.2, because when applying BPE word segmentation, instead of using a maximum number of 90,000 (sub)words sharing the vocabulary between the Basque and Spanish corpora, a separated word segmentation process will be applied for each of the monolingual corpora with a maximum number of 45,000 (sub)words. It is not expected for this error to have significant influence on the results obtained when including the monolingual corpus and its translation, since the automatic evaluation results are very similar when using the aforementioned different BPE segmentations on the corpus used for doing the translation.

4.3 Human evaluation

In this project we will only evaluate the MT results automatically, leaving the necessary human evaluation as a future work. However, in this Section we will briefly describe how this human evaluation could be done, leaving chance to future changes to adapt to the actual systems to evaluate.

In this sense, at first it has to be said that it is expectable that the different approaches described in Section 4.2 will achieve progressively better results as we add subsequent health domain corpora, so the human evaluation method described here should be applied to more diverse systems which obtain high but similar automatic evaluation results, as they can be systems with different architectures or systems that make use of different techniques.

Thus, the evaluation system should compare the at least 3 best different systems that obtain similar automatic evaluation results, taking as a reference around 50 documents containing the diverse kinds of sentences that can be expected in a real scenario in which this kind of systems would be implemented. For the evaluation task, both linguists and health domain experts should be involved, measuring the agreement level between their evaluation results as a way to validate the human evaluation process.

Regarding the linguistic aspects to be evaluated, apart from the commonly used fluency and adequacy measures briefly described in Section 2.1.5, special attention should be paid to the accuracy of the translated terms, for being probably the most important aspect to preserve when developing a MT system for the health domain.

As a complement to this, broader human evaluation processes involving more people could be carried out just asking to choose the best translation from a given set of output sentences generated by the different MT systems to evaluate, providing more reliability to the human evaluation results as the opinion of more people is taken into account.

5 Results

In this chapter we will present the results of the experiments for both parts related to testing different NMT parameters on an out-of-domain corpus (Section 5.1), and evaluating Machine Translation for the health domain with the different systems including succesive health domain corpora in the originally out-of-domain training corpus (Section 5.2). At the end of each subsection we will include a summary to compare the contribution of each of the experiments into the final results.

5.1 NMT parameters test

In this section we will show the obtained results for each of the parameters tested and displayed in Table 6. For a better understanding of the results, in each of the tables shown in this section the first row for each translation direction will correspond to the results obtained with the parameter value selected according to the results shown in the immediately previous section; except for the first table (Optimization), in which the first row for each translation direction will represent the values obtained with the parameters described in Table 5 corresponding to the baseline system.

For each of the tested parameter values and translation directions, BLEU values obtained in dev and test sets will be shown, and training time will also be specified when available (See Section 3.3 for more details). Basque-to-Spanish translation direction will be represented in the tables as 'eu-es', while Spanish-to-Basque will be represented as 'es-eu'.

5.1.1 Optimization

Table 9 shows the results obtained for the 2 tested optimization methods: Adadelta and Adam. In this table we can see that Adam, apart from being significantly faster than Adadelta, obtains better BLEU values for both translation directions and each dataset, with the exception of the test set for Basque-to-Spanish translation direction, in which there is a slight difference of 0.01 points in favour of Adadelta. Since this difference is much higher for Spanish-to-Basque translation direction in favour of Adam (almost 0.3 points) and this trend is also perceived in the dev set, Adam will be chosen as optimization method for succeeding experiments.

Translation direction	**Optimization**	**Training time (hh:mm:ss)**	**dev BLEU**	**test BLEU**
eu-es	Adadelta	213:55:59	26.51	28.98
	Adam	175:24:24	26.87	28.97
es-eu	Adadelta	205:24:57	22.95	20.26
	Adam	187:32:32	23.06	20.55

Table 9: Results for different optimization methods

5.1.2 Unit-type

In Table 10 we can see the results obtained for different tested unit-types: GRU and LSTM. Here we can see that LSTM requires much more training time, even if the showed value for Basque-to-Spanish translation direction can be defined as unusual since it was not reproduced for previous trials in which times more similar to the ones required for Spanish-to-Basque translation direction were needed.

Regarding BLEU values, we can see that GRU obtains better results for Basque-to-Spanish translation direction (same results in the dev set), while LSTM obtains better results for Spanish-to-Basque translation direction. Since the differences in BLEU values are on the edge of being considered significant (0.3 - 0.4 points), and the next experiment consists on testing the beam-width, for which only the evaluation has to be repeated, no decision will be taken for now about which unit-type to choose. Thus, different beam-width values will be tested for both considered unit-types GRU and LSTM.

Translation direction	**Unit-type**	**Training time (hh:mm:ss)**	**dev BLEU**	**test BLEU**
eu-es	GRU	175:24:24	26.87	28.97
	LSTM	265:57:12	26.87	28.68
es-eu	GRU	187:32:32	23.06	20.55
	LSTM	218:04:42	23.37	20.96

Table 10: Results for different unit-types

5.1.3 Beam-width

This section differs from the others because instead of having only one table we will have four tables, one for each fixed value of the parameters unit-type (GRU and LSTM) and beam-width (6 and 10). This way, we will be able to compare easier the different obtained BLEU values and select the optimal values of the tested parameters.

Table 11 shows the results of changing the beam-width for the fixed GRU unit-type. Since beam-width is a parameter related to the evaluation process, we observe that the training time is the same for each translation direction, fact that can be observed also in Table 12.

With regard to BLEU values, we can see that for Basque-to-Spanish translation direction the results are better for a beam-width of 10 for both dev set and test set, while for Spanish-to-Basque translation direction the results are better for a beam-width of 6. However, this difference is a bit higher in Basque-to-Spanish translation direction, with around 0.3 point improvement for beam-width 10, while the advantage for beam-width 6 in Spanish-to-Basque translation direction is equal or lower than 0.2 points for both dev and test sets.

In Table 12 we see the comparison of different beam-width values for the unit-type LSTM. In this case, we can say that better general results are obtained with a beam-width

Translation direction	Beam-width	Training time (hh:mm:ss)	dev BLEU	test BLEU
eu-es	6	175:24:24	26.87	28.97
	10	175:24:24	27.21	29.28
es-eu	6	187:32:32	23.06	20.55
	10	187:32:32	22.92	20.35

Table 11: Results for different beam-width values (unit-type: GRU)

of 10, except for the dev set in Basque-to-Spanish translation direction, where the results are the same as with a beam-width of 6, and the test set in Spanish-to-Basque translation direction, with a slight decrease of 0.03 points.

Translation direction	Beam-width	Training time (hh:mm:ss)	dev BLEU	test BLEU
eu-es	6	265:57:12	26.87	28.68
	10	265:57:12	26.87	28.87
es-eu	6	218:04:42	23.37	20.96
	10	218:04:42	23.64	20.93

Table 12: Results for different beam-width values (unit-type: LSTM)

After comparing the results obtained with different beam-width values for each of the unit-types, now we will compare the results of different unit-types for each fixed values of beam-width 6 and 10. In Table 13 we show the results for beam-width 6, which are the same as shown in Table 10; so we will refer to Section 5.1.2 to analyze the results, only reminding that better results are obtained with GRU for Basque-to-Spanish translation direction, while LSTM outperforms GRU for Spanish-to-Basque translation direction.

Translation direction	Unit-type	Training time (hh:mm:ss)	dev BLEU	test BLEU
eu-es	GRU	175:24:24	26.87	28.97
	LSTM	265:57:12	26.87	28.68
es-eu	GRU	187:32:32	23.06	20.55
	LSTM	218:04:42	23.37	20.96

Table 13: Results for different unit-types (beam-width: 6)

Finally, Table 14 shows the results obtained with a beam-width of 10 for both unit-types GRU and LSTM. Here we can see clearly that GRU obtains better results for Basque-to-Spanish translation direction, while LSTM obtains better results for Spanish-to-Basque translation direction. Furthermore, if we look at all the tables shown in this section, we will see that the best results are obtained for a beam-width of 10, using GRU unit-types

for Basque-to-Spanish translation direction and LSTM for Spanish-to-Basque translation direction; with the only exception of a slight decrease of 0.03 comparing to the results obtained with a beam-width of 6 for Spanish-to-Basque translation direction with LSTM, as stated when commenting the results of Table 12.

Thus, for successive experiments a beam-width of 10 will be used, with GRU unit-types for Basque-to-Spanish translation direction and LSTM unit-types for Spanish-to-Basque translation direction.

Translation direction	**Unit-type**	**Training time (hh:mm:ss)**	**dev BLEU**	**test BLEU**
eu-es	GRU	175:24:24	27.21	29.28
	LSTM	265:57:12	26.87	28.87
es-eu	GRU	187:32:32	22.92	20.35
	LSTM	218:04:42	23.64	20.93

Table 14: Results for different unit-types (beam-width: 10)

5.1.4 Batch-size

In Table 15 we show the results for the different tested batch-size values of 30, 32 and 64. This is the first parameter for which we started to use Mamarro (See Section 3.3), so no training time values are shown for a batch-size of 64. For the other 2 tested values, the required training times are similar for each of the translation directions, with a slight reduction for Basque-to-Spanish translation direction and an even smaller increase for Spanish-to-Basque translation direction (note that different unit-types are used for each translation direction).

Concerning BLEU values, we observe the general trend that increasing the batch-size achieves worse results in the dev set but better results on the test set. This was not observed for previously tested parameters except slight differences smaller than 0.05 points, but since the results on the test set are more relevant than the ones on the dev set, higher values of batch-size will be chosen for succeeding tests. Thus, even if the improvements are lesser than 0.2 points and slightly better results are achieved with a batch-size of 32 for Basque-to-Spanish translation direction, a batch-size of 64 will be used in the following experiments.

5.1.5 Embedding-size

Table 16 shows the results for different tested embedding-sizes of 500, 512 and 1024. First, we observe that the embedding-size of 1024 could not be tested for Basque-to-Spanish translation direction owing to memory restrictions, giving succesive trials an Out Of Memory (OOM) error. In any case, we see that the results on Basque-to-Spanish translation direction clearly decrease when changing the embedding-size from 500 to 512 both for dev

Translation direction	Batch-size	Training time (hh:mm:ss)	dev BLEU	test BLEU
eu-es	30	175:24:24	27.21	29.28
	32	167:58:32	27.08	29.48
	64		27.02	29.45
es-eu	30	218:04:42	23.64	20.93
	32	222:17:53	22.88	20.65
	64		23.05	21.12

Table 15: Results for different batch-size values

and test sets, so it is not expectable that the results could improve for an embedding-size of 1024.

Regarding Spanish-to-Basque translation direction, we see that the results also decrease for higher values of embedding-size on the test set, while a slight improvement of 0.04 points is achieved on the dev set for both embedding-size values of 512 and 1024. Thus, we will consider that this experiment did not led us to any improvement, so an embedding-size of 500 will be selected as optimal among the tested values.

Translation direction	Embedding-size	dev BLEU	test BLEU
eu-es	500	27.02	29.45
	512	26.65	28.87
	1024	OOM	OOM
es-eu	500	23.05	21.12
	512	23.09	20.42
	1024	23.09	20.61

Table 16: Results for different embedding-size values

5.1.6 Comparison with baseline

To sum up this section, a comparison between the results obtained with the initial baseline model and the ones obtained with the optimal values of each of the tested parameters will be carried out. To do so, in Table 17 we show the results of the baseline, characterised by the parameter values described in Table 5, and the best results obtained with each of the parameters tested for both translation directions in both dev and test sets. Note that we do not include the results for embedding-size, since we did not observe any improvement in the conducted experiments; and the results for unit-type correspond to different types GRU and LSTM for each of the translation directions, as the results indicated this was the best option.

When analysing the results on the test set, we observe that the set of experiments carried out results in a 0.47 points increase for Basque-to-Spanish translation direction,

Translation direction	Parameter update	dev BLEU	test BLEU
eu-es	Baseline	26.51	28.98
	Optimization → 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
es-eu	Baseline	22.95	20.26
	Optimization → 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

Table 17: Results for different tested parameters and baseline

and a 0.96 points improvement for Spanish-to-Basque translation direction. In the case of Basque-to-Spanish translation direction, the improvement comes from changing the values of beam-width and batch-size, while for Spanish-to-Basque translation direction the results improved when changing the optimization method, unit-type and batch-size (we can not say that changing the beam-width alone was bad because we did not compare it directly with the baseline but with the updated parameters of optimization method and unit-type).

Therefore, we can conclude that the conducted experiments were mostly satisfactory, except for the embedding-size, and further experiments would be carried out for both beam-width and batch-size. In the case of beam-width this would be faster, since only evaluation would have to be repeated; and regarding batch-size has to be said that the current equipment of Mamarro allows to try higher values as the ones tested in Britz et al. (2017).

5.2 Evaluation on the health domain

In this section we will use the health record models and their manual translations described in Section 3.2.3 to firstly evaluate the optimal system among the ones tested in the previous section on the health domain, and then subsequent health domain corpora will be added to the out-of-domain training corpus to measure their influence on the translation task.

As in previous section, each of the experiments will be carried out for both translation directions Basque-to-Spanish ('eu-es') and Spanish-to-Basque ('es-eu'), with the aforementioned exception of including the monolingual corpus and its translation, that will only be tested for Basque-to-Spanish translation direction. As before, BLEU values obtained in dev and test sets will be shown for each of the conducted experiments. In this case, all the experiments will be carried out using Mamarro (See Section 3.3 for more details), so no information about training time will be provided.

5.2.1 Using the out-of-domain corpus

Table 18 shows the results of the best system trained in the previous section with an out-of-domain corpus, but in this case using a health domain corpus for evaluation.

Translation direction	dev BLEU	test BLEU
eu-es	10.69	10.67
es-eu	9.08	8.69

Table 18: Results on the health domain with the out-of-domain corpus

As expected when using different domain corpora for training and evaluation, the results are very poor, so these will be interpreted just as a reference to see how much each of the following experiments contribute to a better translation on the health domain.

5.2.2 Including a health-related dictionary

For the first of these experiments, a dictionary built with the SNOMED CT Spanish terms and the corresponding automatically created Basque terms will be included to the out-of-domain corpus. Table 19 shows the results for this configuration.

Translation direction	dev BLEU	test BLEU
eu-es	15.45	15.04
es-eu	10.75	10.44

Table 19: Results on the health domain including a health-related dictionary

In this case, we observe that the results have significantly improved for Basque-to-Spanish translation direction in both dev and test sets, reaching an almost 4.4 points gain in the test set comparing to the results obtained using only an out-of-domain corpus for training. Regarding Spanish-to-Basque translation direction, we achieve a gain of 1.7 points in both dev and test sets, even if the results are still low. However, we remark that the influence of the health-related dictionary has proved to be effective for both translation directions, despite the big difference between the sizes of the original corpus (4.5M sentences) and the added health domain dictionary (151,111 entries).

5.2.3 Including artificial sentences created from SNOMED CT

Regarding the inclusion of artificial sentences created from the relational information stored in SNOMED CT, Table 20 shows the results after including these sentences to the previous training corpus containing the out-of-domain corpus and the health domain dictionary.

Comparing with the results obtained only adding the health domain dictionary, we observe that the results have slightly improved for Basque-to-Spanish translation direction, gaining 0.6 BLEU points in the dev set and 0.4 points in the test set, while the results have remained almost invariable for Spanish-to-Basque translation direction.

Translation direction	dev BLEU	test BLEU
eu-es	16.08	15.48
es-eu	10.79	10.43

Table 20: Results on the health domain including artificial sentences created from SNOMED CT

Looking for the possible reasons of this unexpectedly low or even zero improvement when adding the artificial sentences, we have to point out that this inclusion did not suppose any enrichment from the morphological perspective, since all the medical terms had already been added when adding the dictionary; and from the syntactic point of view, the results indicate that the defined sentence models, created according to the relations on SNOMED CT, did not reflect the characteristic syntax of the health record models used for evaluation.

5.2.4 Including a monolingual corpus and its translation

Finally, Table 21 shows the results for the Basque-to-Spanish translation direction after including in the training corpus the Spanish monolingual corpus and its machine translation in Basque.

Translation direction	dev BLEU	test BLEU
eu-es	22.52	21.07

Table 21: Results on the health domain including a monolingual corpus and its translation

In this case, we observe that the application of the backtranslation technique has been greatly beneficial comparing to the previous systems including the dictionary and the sentences created from SNOMED CT, improving the results up to 6.4 BLEU points in the dev set and 5.6 points in the test set.

As a reference, in the original paper describing the backtranslation technique (Sennrich et al., 2015a), improvements of around 3 BLEU points are reported; however, we have to state that in this case our baseline system obtains much lower results and, on the other hand, the added corpus is also helpful for domain adaptation to the health related documents used for evaluation.

Regarding the size of the corpus used for backtranslation, a recent study showed that the inclusion of more and more automatically translated data could be helpful as long as it does not exceed the double of the size of the original bilingual corpus (Poncelas et al., 2018). In our case, the available Spanish monolingual corpus is formed by less than half of the number of sentences from the bilingual out-of-domain corpus (2,023,811 sentences for backtranslation added to 4,530,683 sentences from the out-of-domain corpus), which indicates that there is still room from improvement in case that more monolingual corpora becomes available.

5.2.5 Summary of results on the health domain

For measuring the influence of each of the experiments carried out in this section, Table 22 groups the results obtained when evaluating on a health domain corpus systems using different corpora for training. As all the health domain corpora were successively added to the out-of-domain corpus, '+' sign should be interpreted as an addition to the corpus corresponding to the immediately upper row.

Translation direction	**Training corpus**	**dev BLEU**	**test BLEU**
eu-es	out-of-domain	10.69	10.67
	+ dictionary	15.45	15.04
	+ art. sentences	16.08	15.48
	+ backtranslation	22.52	21.07
es-eu	out-of-domain	9.08	8.69
	+ dictionary	10.75	10.44
	+ art. sentences	10.79	10.43

Table 22: Results on the health domain with different training corpora

Analysing the results globally, we observe that all the conducted experiments have improved the results to a greater or lesser degree, except the inclusion of artificial sentences that has not proved to be beneficial for Spanish-to-Basque translation direction.

Regarding the different translation directions, we observe that the inclusion of each of the health-related corpora has been more useful for Basque-to-Spanish translation direction, specially for the system including only the dictionary, where a 4.4 BLEU points gain was achieved in the test set for Basque-to-Spanish translation direction comparing to a 1.7 points gain in the same set for Spanish-to-Basque translation direction.

Finally, examining the results of including the different health domain corpora, 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 showed 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.

For future experiments, we have to point out that even if bilingual health domain corpora would be 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.

Translation examples

Before presenting the conclusions of this project, we will show some of the translations carried out by the different tested systems of selected sentences from the dev and test sets. Table 23 and 24 show two sentences from the dev set and another two from the test set respectively, along with the translations generated by the systems trained with increasing presence of health domain corpora for Basque-to-Spanish translation direction.

Analysing the sentences translated by different systems, we observe that the system trained only with the out-of-domain corpus is unable to translate some basic medical terms like 'reflujo', 'glaucoma' or even 'ojo' (all in the first sentence selected from the dev set), giving a strange translation for part of the second sentence extracted from the test set ('tendinitis de hombro': 'la avioneta de sorbalde').

Regarding the system including the dictionary, we observe a great improvement in terms of the vocabulary that is able to translate, as we can see with the aforementioned terms from the first sentence selected from the dev set, or even 'antibiótico' in the second sentence selected from the dev set; but it is still uncapable of translating other terms like 'maleolo' in the first sentence extracted from the test set, or 'hombro', translated like 'césped' in the second sentence extracted from the test set.

With respect to the system including the artificial sentences created from SNOMED CT, we only observe little improvements comparing to the system including the dictionaries, as we can see in the first sentence selected from the dev set, where the system outputs 'dolores de cuello' instead of the previous 'dolores collares'; or in the same examples from the test set mentioned in the previous paragraph ('maléolo' and 'hombros', in this case with an unnecesary plural suffix and associated with 'neumonía' instead of 'tendinitis').

Finally, we can say that the system including the monolingual corpus and its translation gives the best results not only in terms of the specific vocabulary that is able to translate, as we can see with the term 'intravenoso' in the second sentence selected from the dev set or 'palpación' in the first sentence extracted from the test set; but also in terms of syntax, as we can see in part of the second sentence extracted from the test set, being the only system that translates correctly 'tendinitis de hombro y neumonía'.

Language Analysis and Processing

Original sentence in Basque	beste batzuk : errefluxu gastroesofagikoa ; lepoaldeko minak ; glaukoma / ezkerreko begiko katarata
Manual translation into Spanish	otras : reflujo gastroesofágico , cervicalgia , glaucoma / catarata izqda.
Translation by the system trained with the out-of-domain corpus	otros : repliegue gastroesofágico ; minas favorables ; glares ma / catarata de vegas de izquierda
Translation by the system trained including a health-related dictionary	otros : reflujo gastroesofágico ; dolores collares ; glaucoma / catarata del ojo izquierdo
Translation by the system trained including artificial sentences created from SNOMED CT	otros : reflujo gastroesofágico ; dolores de cuello , glaucoma / cataratas del ojo izquierdo
Translation by the system trained including a monolingual corpus and its translation	otros : reflujo gastroesofágico ; dolores a nivel cervical ; glaucoma / catarata de ojo izquierdo
Original sentence in Basque	pazientea ospitaleratu egin dugu zain barneko tratamendu antibiotikoa egiteko , eta ez du konplikaziorik izan eboluzioan
Manual translation into Spanish	se le hospitaliza para seguir tratamiento antibiótico y no ha tenido complicaciones en su evolución
Translation by the system trained with the out-of-domain corpus	la paciente se encuentra ingresada en un hospital para realizar un tratamiento de tratamiento externo , y no tiene complicaciones en la evolución
Translation by the system trained including a health-related dictionary	el paciente ha ingresado en el hospital para realizar un tratamiento antibiótico interno y no ha tenido complicaciones en la evolución
Translation by the system trained including artificial sentences created from SNOMED CT	el paciente , que ha sido hospitalizado para realizar un tratamiento antibiótico interno , no ha sufrido complicaciones en la evolución
Translation by the system trained including a monolingual corpus and its translation	la paciente es ingresada para realización de tratamiento antibiótico intravenoso y no ha tenido complicaciones durante la evolución

Table 23: Sample of sentences from the dev set along with the output of the different tested systems for Basque-to-Spanish translation direction

Original sentence in Basque	kanpoko maleoloaren haztapen mingarria , baita barne-maleoloarena ere
Manual translation into Spanish	palpación dolorosa de maléolo externo , así como el interno
Translation by the system trained with the out-of-domain corpus	hiriente doloroso del maleolio exterior , y también del llamado maleador interno
Translation by the system trained including a health-related dictionary	aumento dolorosa de la maleolia externa , así como de la maleóloga interna
Translation by the system trained including artificial sentences created from SNOMED CT	flexión doloroso del maléolo externo , incluso del maléolo interno
Translation by the system trained including a monolingual corpus and its translation	palpación dolorosa del maleolo externo así como del maleolo interno
Original sentence in Basque	sorbaldetako tendinitisa eta pneumonia , 2012ko otsailean
Manual translation into Spanish	en febrero de 2012 tendinitis de hombro y neumonía
Translation by the system trained with the out-of-domain corpus	la avioneta de sorbalde y la neumonía , en febrero de 2012
Translation by the system trained including a health-related dictionary	una tendinitis y neumonía de césped en febrero de 2012
Translation by the system trained including artificial sentences created from SNOMED CT	tendinitis y neumonía de los hombros , en febrero de 2012
Translation by the system trained including a monolingual corpus and its translation	tendinitis de hombro y neumonía en febrero de 2012

Table 24: Sample of sentences from the test set along with the output of the different tested systems for Basque-to-Spanish translation direction

6 Conclusions and future work

In this section we will present the conclusions drawn from the results obtained in the different experiments (Section 6.1), and we will end mentioning the future work that could be carried out in this area (Section 6.2):

6.1 Conclusions

Conditioned by the lack of bilingual corpora for the health domain in Basque and Spanish languages, the conclusions of this project will be divided in the same way as the developed experiments, taking on the one hand the results of the evaluation of different NMT parameters tested on systems trained with an out-of-domain corpus; and on the other hand, the results of evaluating with health domain texts the systems trained with different corpora with increasing presence of health domain corpora.

Regarding the use of different NMT parameters, we conclude that the conducted experiments have been positive in general, with almost 0.5 points gain in BLEU for Basque-to-Spanish translation direction, and almost 1 point improvement for Spanish-to-Basque translation direction comparing to the results obtained with an already strong baseline. In particular, we observe that the use of Adam as optimization method and LSTM as unit-type has been beneficial for Spanish-to-Basque translation direction, while the increase of beam-width to 10 has improved the results for Basque-to-Spanish translation direction and the augmentation of batch-size to 64 has proved to be positive for both translation directions.

In the case of adding different health domain corpora for evaluation on the health domain, we remark the great improvement achieved through the technique of backtranslation, achieving a 5.6 BLEU points gain for the tested Basque-to-Spanish translation direction. We also observe that the inclusion of the health-related dictionary has significantly improved the results, specially for Basque-to-Spanish translation direction, obtaining a 4.4 BLEU points gain compared to a 1.7 points gain for Spanish-to-Basque translation direction. Altogether, the applied improvements have made possible to obtain an acceptable result of 21.07 BLEU points for Basque-to-Spanish translation direction, even without using bilingual health domain corpora.

However, we have to state that all the above conclusions are based on automatic evaluation metrics, which we know that not always reflect perfectly the quality of a given MT system. Therefore, before developing a real system for Basque/Spanish Machine Translation for the health domain, a human evaluation process must be carried out to test the quality of the developed systems.

6.2 Future work

1) The first needed contribution to be able to develop a Basque/Spanish Machine Translation system for the health domain would be to collect a big enough corpus containing EHRs in Basque and Spanish. This project is already underway with the Basque public health service (Osakidetza), and is expectable that it will be available soon for future research in this area.

2) Once this health domain bilingual corpus would be available, NMT systems as the ones tried in Section 5.1 could be trained and tested directly with a health domain corpus.

3) Apart from this, as stated in Section 5.1.6, further experiments could be carried out with higher values of beam-width and batch-size.

4) Meanwhile, other NMT architectures like the recently described Transformer (Vaswani et al., 2017) could be tested, specially taken into account that they obtain better results than other architectures when the available corpus is reduced.

5) Adding to the experiments done to test the effects of including a health domain dictionary or a monolingual corpus to the out-of-domain bilingual corpus, more sophisticated techniques like the ones expressed in Gu et al. (2017) and Zhang et al. (2018) could be explored. These works extend the idea of NMT by using semantic similarity to search sentences similar to the ones to translate and include them in the training corpus (Gu et al., 2017) or use them to rescore the output probabilities of NMT systems (Zhang et al., 2018), achieving BLEU improvements up to 6 points with legal domain corpora.

6) In addition to the different NMT settings that could be tried, already existing RBMT and SMT systems for Basque language could be adapted and tested for the health domain. To do so, English-to-Basque RBMT system for the health domain MatxinMed (Perez-de-Viñaspre, 2017) should be adapted to Basque/Spanish language pair; and already existing EuSMT (Labaka, 2010) should be trained with the expected health domain corpus in Basque/Spanish.

7) Finally, MT technique hybridization could be tried, for instance, using the adapted RBMT and SMT systems to translate a Spanish monolingual health domain corpus to Basque and include the input and output sentences as training corpus for NMT systems. This way, linguistic information from RBMT systems and statistical information from SMT systems could be exploited by NMT systems.

8) For all of the possible systems described above or any other that could be tested in the future, a human evaluation like the one designed in Section 4.3 should be done to have a clearer comparison of the different developed systems.

References

Eneko Agirre, Inaki Alegria, Xabier Arregi, Xabier Artola, A Díaz de Ilarraza, Montse Maritxalar, Kepa Sarasola, and Miriam Urkia. Xuxen: A spelling checker/corrector for basque based on two-level morphology. In *Proceedings of the third conference on Applied natural language processing*, pages 119–125. Association for Computational Linguistics, 1992.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. Findings of the 2016 conference on machine translation. In *ACL 2016 FIRST CONFERENCE ON MACHINE TRANSLATION (WMT16)*, pages 131–198. The Association for Computational Linguistics, 2016.

Denny Britz, Anna Goldie, Thang Luong, and Quoc Le. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*, 2017.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

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

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

Marta R. Costa-Jussà and José A. R. Fonollosa. Character-based neural machine translation. *arXiv preprint arXiv:1603.00810*, 2016.

Hatem M Elsherif and Tariq Rahim Soomro. Perspectives of arabic machine translation. *Journal of Engineering Science and Technology*, 12(9):2315–2332, 2017.

Thierry Etchegoyhen and Gorka Labaka. Modela project. NMT workshop. Donostia, Euskal Herria., 2017.

Thierry Etchegoyhen, Eva Martínez Garcia, Andoni Azpeitia, Gorka Labaka, Iñaki Alegria, Itziar Cortes Etxabe, Amaia Jauregi Carrera, Igor Ellakuria Santos, Maite Martin, and

Eusebi Calonge. Neural machine translation of basque. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 139–148, 2018.

Mikel L Forcada and Ramón P Ñeco. Recursive hetero-associative memories for translation. In *International Work-Conference on Artificial Neural Networks*, pages 453–462. Springer, 1997.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. Search engine guided non-parametric neural machine translation. *arXiv preprint arXiv:1705.07267*, 2017.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*, 2017.

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

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

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

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

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

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 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*, pages 48–54. Association for Computational Linguistics, 2003.

Gorka Labaka. *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., 2010.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

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

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

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

Peyman Passban, Qun Liu, and Andy Way. Improving character-based decoding using target-side morphological information for neural machine translation. *arXiv preprint arXiv:1804.06506*, 2018.

Olatz Perez-de-Viñaspre. *Osasun-alorreko termino-sorkuntza automatikoa: SNOMED CTren eduki terminologikoaren euskaratzea*. PhD thesis, University of the Basque Country, Donostia, Euskal Herria., 2017.

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

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

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

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*, 2016.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

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

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Coverage-based neural machine translation. 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

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

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. Guiding neural machine translation with retrieved translation pieces. *arXiv preprint arXiv:1804.02559*, 2018.

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