Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 121-132 recibido 11-12-2020 revisado 25-01-2021 aceptado 03-02-2021

Grammatical error correction for Spanish health records

Corrección de errores gramaticales en informes clínicos en español

Salvador Lima-López,[1,*] **Naiara Perez,**[2] **Montse Cuadros**[2]
[1]Barcelona Supercomputing Center, Barcelona, Spain
[2]SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Mikeletegi Pasealekua 57, Donostia/San-Sebastián, 20009, Spain
salvador.limalopez@bsc.es, {nperez, mcuadros}@vicomtech.org

Abstract: This paper describes the first approach to Grammatical Error Correction for Spanish health records. We present a series of experiments using neural networks and data augmentation, achieving 70.89 $F_{0.5}$ score. Resources designed for this task are introduced, namely the IMEC corpus of corrected health records and the TMAE corpus of clinical texts augmented with errors.
Keywords: health records, Grammatical Error Correction, Spanish.

Resumen: Este artículo presenta el primer trabajo sobre la corrección gramatical de textos clínicos en español. En este trabajo, presentamos un conjunto de experimentos basados en redes neuronales y aumentación de datos, en los cuales conseguimos una puntuación de 70,89 $F_{0,5}$. Además, se presentan dos corpus creados para esta tarea: el corpus IMEC, un corpus médico corregido manualmente, y el corpus TMAE, un corpus de textos clínicos aumentado con errores.
Palabras clave: informes clínicos, corrección de errores gramaticales, español.

1 Introduction

Grammatical Error Correction (GEC) is a field within Natural Language Processing that deals with the correction of texts from a grammatical, lexical and orthographic point of view. As a discipline, it has traditionally focused on educational applications such as second-language learners' essays. However, there are other text genres that can also benefit from this sort of treatment. One of them is health records, a type of clinical text. Health records are documents where doctors describe patients' consultations to their office, including their impressions, diagnosis and recommendations.

As the main source of written communication between health professionals and patients, as well as among health specialists themselves, health records should be written in the most correct way possible. Still, due to the heavy time restrictions that health professionals usually work under, health records often present strange grammar structures, abbreviated words and outright spelling errors. Consider the example in Table 1. These documents sometimes end up being the source of misunderstandings on the patients' side (Terroba Reinares, 2015, p. 11).

Original:
EEII: no edemas ni singos de tvp.
Corrected:
EEII: no muestra edemas ni signos de TVP.
Translation:
LE: no signs of edemas nor TVP.

Table 1: A real sentence extracted from a health report verbatim and the proposed correction.

Given that in Spain it is legally required for doctors to write a health record for each consultation (Boletín Oficial del Estado, 2015) and that, according to the latest data available, in 2018 there were over 350 million Primary Health Care and nursing consultations (Ministerio de Sanidad, 2018, p. 11), it is safe to assume that it is not feasible to manually revise and correct health records.

In this paper, we explore for the first time

*Work done while at Vicomtech.

ISSN 1989-7553. DOI 10.26342/2021-66-10

the feasibility of making health records in Spanish clearer and more accessible by applying GEC techniques. We obtain promising results using two new corpora specifically curated for this task. These corpora are, on the one hand, a collection of manually corrected Spanish health records named IMEC (*Informes Médicos en Español Corregidos*, or Corrected Health Records in Spanish) and, on the other, a compilation of various clinical corpora artificially augmented with errors called TMAE (*Textos Médicos Aumentados con Errores* or Clinical Texts Augmented with Errors).

The structure of this paper is the following: §2 briefly discusses the history and current developments of GEC; §3 explains two different corpora for clinical GEC, while §4 introduces the different experiments performed with them and discusses their results. §5 concludes this work by making some final remarks and discussing possibilities of future work.

2 Related Work

Early GEC systems date back to the 1980s, where rule-based pattern recognisers and dictionary-based systems (Macdonald, 1983; Richardson and Braden-Harder, 1988) were initially used. Later on, statistical classifiers were also implemented, focusing on specific error types (Gamon et al., 2008; Tetreault, Foster, and Chodorow, 2010; Lee and Seneff, 2008).

The most successful approach has been to treat GEC as a Machine Translation (MT) task. An analogy can be drawn between both fields, where MT's source language corresponds to GEC's uncorrected text and the target language to corrected text. Statistical MT systems made possible the generation of an N-best list of alternative corrections for each sentence (Shen, Sarkar, and Och, 2004), which can be re-ranked using text features, classifiers or language models. Reranking helps improve overall performance and has become a staple of many state-of-the-art GEC systems even nowadays.

Neural networks have also been proposed due to their generalization potential. For a long time, the most popular architecture has been the Encoder-Decoder model, accompanied either by recurrent neural networks (Xie et al., 2016) or convolutional neural networks (Chollampatt and Ng, 2018). Lately, in the same fashion as many other NLP tasks, the state of the art has been achieved using Transformers (Vaswani et al., 2017).

It could be argued that shared tasks have played an important role in the development of this sub-field. The Conference on Natural Language Learning (CoNLL) held a shared task on GEC both in 2013 (Ng et al., 2013) and 2014 (Ng et al., 2014), releasing a different corpus each year. In 2019, the Building Educational Applications (BEA) shared task was held (Bryant et al., 2019). It saw the release of new annotated datasets (W&I (Yannakoudakis et al., 2018)+LOCNESS (Granger, 1998)), as well as the re-release of previously available corpora (FCE (Yannakoudakis, Briscoe, and Medlock, 2011), LANG-8 (Tajiri, Komachi, and Matsumoto, 2012; Mizumoto et al., 2012) and NUCLE (Dahlmeier, Ng, and Wu, 2013)) in a standardized version. This process was performed using ERRANT (Felice, Bryant, and Briscoe, 2016), a toolkit specifically designed for the annotation of GEC data. At the time of this writing, the state of the art for CoNLL 2014's dataset and W&I+LOCNESS is a Transformer model called GECToR (Omelianchuk et al., 2020) that is trained with sentences augmented with artificial errors and fine-tuned on real data.

Using artificial errors, as GECToR does, is a common technique in GEC known as Artificial Error Generation (AEG). It consists in introducing errors into error-free sentences to create parallel correct/incorrect pairs. While GECTor learns how to make these changes using Machine Learning, rule-based systems can also be used for this task. For instance, Beloki et al. (2020) created a parallel GEC corpus of 500,000 news in Basque using grammatical rules. AEG is a technique that can be performed on any text of any genre and that is very flexible since both the type of errors and how they are introduced (i.e., randomly or probabilistically) can be controlled. This method was studied in-depth, among others, by Felice (2016), Rei et al. (2017) and Grundkiewicz, Junczys-Dowmunt, and Heafield (2019).

Along the same lines, oversampling is a technique that is usually applied to balance unbalanced corpora in classification tasks. It is implemented in some low-resource GEC research due to its seemingly good results

(Náplava and Straka, 2019). In low-resource GEC, however, all sentences are duplicated regardless of whether the errors they include are from a minority class or not, as the objective is not to balance the corpus but simply to expand it. All in all, data augmentation is often a part of GEC due to the sparsity of quality parallel data.

So far, most research on Grammatical Error Correction has focused on English texts. In Spanish, the Corpus Of Written Spanish–L2 and Heritage speakers (COWS-L2H) (Davidson et al., 2020) was recently released and its authors tested its validity by training a GEC system based on an LSTM (Long Short-Term Memory) encoder-decoder. There does not seem to be any other recent GEC papers focused on Spanish. Regarding clinical texts, to the best of our knowledge there are no studies that apply GEC techniques to this domain.

3 Corpora

This section presents the two corpora developed for GEC in the clinical domain: IMEC (*Informes Médicos en Español Corregidos*, or Corrected Health Records in Spanish) and TMAE (*Textos Médicos Aumentados con Errores* or Clinical Texts Augmented with Errors).

3.1 Corrected Health Records in Spanish (IMEC)

IMEC is a collection of sentences from Electronic Health Records presented as parallel correct/incorrect sentence pairs. The sentences have been extracted from NUBes (Lima López et al., 2020), a corpus of Electronic Health Record in Spanish manually labelled with negation and uncertainty phenomena. A sample sentence taken from IMEC is shown in Table 2:

Original:
En Abril de 2003 en escreenning de cancer [...]
Corrected:
En abril de 2003 en un screening de cáncer [...]
Translation:
On April 2003, in a cancer screening [...]

Table 2: A original-corrected sentence pair from IMEC.

The corpus consists of 10,007 sentences, of which 7,801 have at least one correction. The sentences were manually corrected by a single annotator.[1] Correction guidelines were developed based on two different style guides as reference: Bello Gutiérrez (2016) and Aguilar Ruíz (2013). The principles underlying the guidelines are three:

(i) terminological and semantic errors are not be considered, as they should only be corrected by health professionals;

(ii) even if abbreviations are one of the main sources of ambiguity in clinical texts, the only changes made to them is to normalize their spelling; abbreviation disambiguation is in its own a NLP task of great difficulty, particularly in the health domain; and,

(iii) a text's clarity comes from both its content and its structure, which means that our corrections should cover orthotypographic (spelling, punctuation) as well as grammatical aspects.

The parallel corrected/uncorrected sentences have been annotated using the ERror ANnotation Toolkit (ERRANT) (Felice, Bryant, and Briscoe, 2016). ERRANT aligns parallel sentences, extracts edits and categorizes them according to a bi-axial system. The first axis corresponds to the types of changes made to the text:

- Missing: the correction consists in inserting a missing token in the incorrect sentence.
- Unnecessary: the correction consists in deleting a token from the incorrect sentence.
- Replacement: the correction consists in replacing a token with another.

The second axis classifies the errors by the linguistic properties of the token(s) involved. This classification is carried out by ERRANT with rules that involve part-of-speech tagging with the Universal Dependency tagset (Bryant, 2019). Some of the error types include NOUN for noun-related errors, DET for determiner-related errors, SVA for subject-verb agreement errors, and so on.

[1] Even though having a corpus corrected by only one annotator is not ideal, reannotating the corpus in the future is a possibility. For instance, the CoNLL-2014 test set was reannotated multiple times, up to a total of 18 overlapping annotations (Bryant, 2019).

Type	#	%
Replacement	14,414	53.20
Missing	12,505	46.13
Unnecessary	184	0.67
SPELL	12,024	44.36
DET	6,959	25.68
PUNCT	3,829	14.13
PREP	1,254	4.63
VERB	921	3.40
OTHER	787	2.90
ORTH	388	1.43
NOUN	367	1.35
MORPH	81	0.30
AUX	76	0.28
ADJ:INFL	66	0.24
NOUN:INFL	65	0.24
CONJ	57	0.21
DET:INFL	56	0.21
ADJ	42	0.15
PRON	37	0.14
VERB:SVA	25	0.09
VERB:TENSE	24	0.09
VERB:FORM	19	0.07
ADV	15	0.06
SCONJ	7	0.03
WO	4	0.01

(a) IMEC

Type	#	%
Replacement	4,728,619	60.66
Missing	3,021,303	38.76
Unnecessary	44,099	0.56
SPELL	2,387,712	30.64
DET	2,017,786	25.89
PUNCT	1,394,900	17.90
PREP	1,146,550	14.71
OTHER	322,858	4.14
VERB	236,658	3.04
DET:INFL	95,764	1.23
NOUN	50,418	0.65
MORPH	39,468	0.51
ORTH	26,665	0.34
CONJ	23,973	0.31
ADJ:INFL	10,845	0.14
PRON	9,745	0.13
ADJ	9,521	0.12
NOUN:INFL	7,668	0.10
ADV	7,174	0.09
VERB:TENSE	3,437	0.04
AUX	1,420	0.02
SCONJ	933	0.01
WO	352	0.00
VERB:FORM	197	0.00
VERB:SVA	2	0.00

(b) TMAE

Table 3: Edit and error type distribution in the corpora IMEC and TMAE.

More general types also exist, such as SPELL for spelling errors or OTHER for edits that do not fit into any other category.

ERRANT's rules and resources are originally designed for English but we have adapted them to Spanish for the annotation of IMEC. Some of the changes to the rules include:

- The category ADJ:FORM, renamed to ADJ:INFL, includes gender and number agreement errors.
- New category called DET:INFL added for determiner-noun agreement errors.
- NOUN:INFL now encompasses all noun agreement errors; NOUN:NUM is deprecated.
- NOUN:POSS category was eliminated as there is not possessive inflection for nouns in Spanish.
- Addition of new rules for specific spelling (SPELL) mistakes (e.g., accentuation).

The distribution of corrections in IMEC, both in terms of edit type and of error type, is shown in Table 3a. Regarding edit types, the most common are replacements, followed by missing tokens. Unnecessary tokens are rare, partly due to the annotation guidelines mentioned above.

When it comes to error types, there is a clear unbalance in the corpus. Most of the corrections are concentrated on the orthotypographic aspects, mainly spelling and punctuation. Grammar errors are not as common as in other GEC corpora such as FCE (Yannakoudakis, Briscoe, and Medlock, 2011) or NUCLE (Dahlmeier, Ng, and Wu, 2013). These corpora usually contain texts written by language learners. In contrast, IMEC's original authors are native speakers. Their grammar is usually correct but they are less careful in other aspects. Even then, there are many errors related to determiners, prepositions and verbs due to the health professionals' style being quite telegraphic.

Corpus	# Lines	# Tokens
IBECS	1,035,660	25,157,063
SciELO	919,553	26,706,151
PubMED	354,724	4,558,980
SPACCC	15,907	416,494
TMAE incorrect	2,325,844	51,710,613
TMAE correct	2,325,844	56,841,053

Table 4: Size of TMAE and its constituents in terms of the total number of lines and the total number of running tokens.

3.2 Clinical Texts Augmented with Errors (TMAE)

TMAE is the result of a merger of health-related texts from various sources that were pre-processed and induced with errors in order to create a synthetic corpus for GEC. Four different corpora were chosen to be augmentated: IBECS, SciELO, Pubmed (all three are part of the MeSpEn collection by Villegas et al. (2018))[2] and SPACCC (Intxaurrondo, 2018). Altogether, the resulting corpus has a size of over 2.3 million parallel sentences and 51 million tokens with almost 8 million annotations. Table 4 shows in detail the sizes of the different corpora used in TMAE. "TMAE correct" stands for the merger of the different corpora.

The aim of the augmentation was to introduce errors in a way that replicated the error types and distribution found in the IMEC corpus. For this purpose, a set of rules was handcrafted that recreated the most prominent errors in the corpus. The changes include adding or removing words based on their part-of-speech tag, introducing typos or changing the inflection of a word. A total of 24 different rules were developed, each with an assigned probability based on the frequency in IMEC of the error they generate.

The number of edits a sentence experiences is randomly chosen between 1 and 4. To avoid completely changing a sentence, sentence length was also taken into account to set a maximum threshold of edits. This corpus with introduced errors is "TMAE incorrect" in Table 4. The examples in Table 5 show some of the resulting sentence pairs.

As with IMEC, once the parallel sentences were generated, the whole corpus was annotated using ERRANT. Table 3b describes the error distribution of the corpus. When compared with IMEC, most categories are similarly distributed, although some such as PREP or OTHER have grown, and others such as SPELL have decreased in size.

Original:
Aplicación de la metodología enfermera en pacientes con úlceras por presión.
Augmented:
Aplicacion metodología enfermera pacientes con úlceras por presión
Translation:
Application of the nursing methodology in patients with pressure ulcers
Original:
También se discute la necesidad de controlar con imagen la resolución de la TEP tras el tratamiento anticoagulante, actualmente no recomendado en las guías clínicas.
Augmented:
También se discute necesidad de controlar con imagen resolución de la TEPtras el tratamiento anticoagulante, actual no recomendado en las guías clínicas.
Translation:
The need for imaging the resolution of PE after anticoagulant treatment is also discussed, currently not recommended in clinical guidelines.

Table 5: Examples of automatically induced errors.

4 Experimentation

This section documents the experimentation details followed using the resources presented above, from the development of a baseline to training a neural network. Next, we explain the experimentation setup. Results are presented in §4.2 and discussed in §4.3.

4.1 Experimentation design

Figure 1 shows the workflow of our experimentation. IMEC provides training and development data, as well as the gold standard against which to measure the results of the experiments. The sizes of these partitions are shown in Table 6. We also rely on TMAE in order to increase the volume of the training data artificially.

Two systems are evaluated: Aspell,[3] which sets the baseline, and a Multilayer Convolutional Encoder-Decoder (Chollampatt and Ng, 2018) as a more sophisticated solution. Each of these systems is extensively

[2] https://temu.bsc.es/mespen/

[3] https://aspell.net

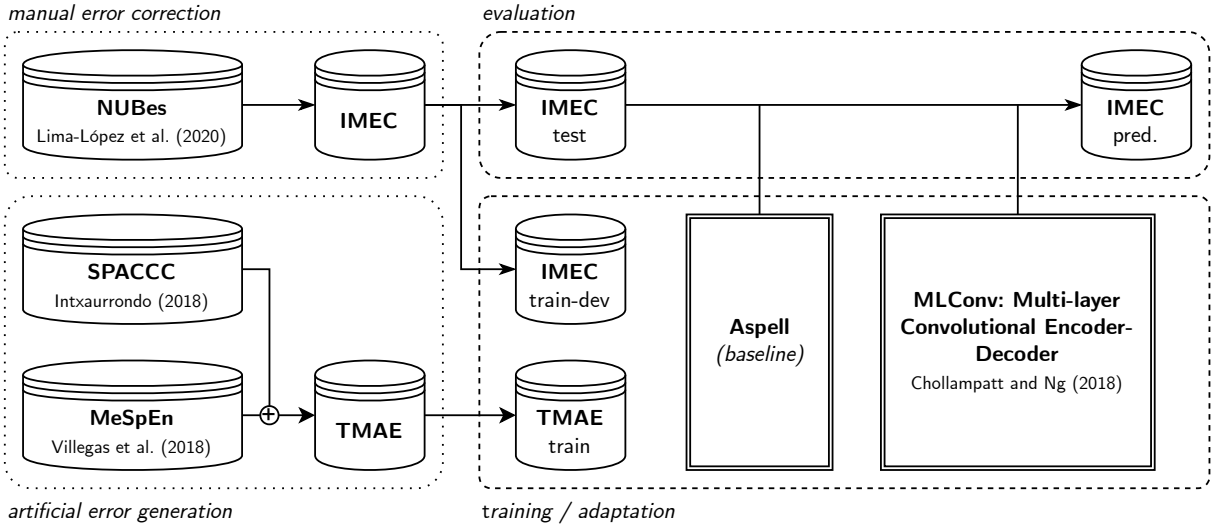

Figure 1: Overview of the different stages of this work.

Partition	# Lines	%
Train	7,507	75
Dev	1,500	15
Test	1,000	10

Table 6: Partitions' size of the IMEC corpus.

evaluated using different key-component settings and/or training set variations.

The output of each system for the test set has been annotated and scored using ERRANT. ERRANT evaluates performance in terms of the $F_{0.5}$ measure, which weights precision twice as much as recall.

4.1.1 Baseline

A baseline system was created as a benchmark for the experimentation. Given the number of spelling mistakes in the IMEC corpus, it was decided that a spell checker would be enough for the task even if it wasn't able to tackle all error types. Ultimately, the free software Aspell was chosen, as it is a renowned spellchecker that allows for some customization.

Aspell uses a dictionary to check whether a word is correctly spelled and suggests possible replacements for misspelled terms based on the metaphone algorithm and a variant of the Levenshtein distance (Atkinson, 2020).

Some of the customizations allowed include using custom dictionaries or applying filters. It is able to return a list with all possible suggestions, which can be further processed. Thus, we tweaked out-of-the-box Aspell as follows:

(i) the predetermined Spanish dictionary was expanded using a vocabulary list extracted from IMEC's train set;

(ii) a Levenshtein distance threshold for suggestions was set; and,

(iii) the suggestions provided were re-ranked using a language model.

The language model was trained on the MedlinePlus corpus, which is part of the MeSpEn collection (Villegas et al., 2018), using the KenLM toolkit (Heafield, 2011) with a window size of 5. Its size can be consulted in Table 7.

4.1.2 MLConv

As a competitive system, we performed experiments with Chollampatt and Ng (2018)'s multilayer convolutional encoder-decoder, MLConv.[4] It is a model that consists of an encoder convolutional network followed by a decoder convolutional network, each with seven layers.

The input to the network are fastText (Bojanowski et al., 2017) word embeddings pre-trained on data segmented with the byte-pair encoding (BPE) algorithm (Sennrich, Haddow, and Birch, 2016). BPE splits rare words into sub-words, helping minimize the number of out-of-vocabulary words. The output of the decoder is an N-best list of corrections. Each candidate is re-ranked using a log-linear framework that calculates features weights on the development set using min-

[4] https://github.com/nusnlp/mlconvgec2018

inum error rate training (Och, 2003). Additionally, edit operation and language model features are also used for scoring.

In this work, the fastText word embeddings were trained on a 2018 dump of the Spanish Wikipedia.[5]

Furthermore, three language models were tested to study the effect of in- and out-of-domain knowledge in the re-ranking step:

- MedlinePlus: the same medical corpus used for the Aspell baseline, MedlinePlus (Villegas et al., 2018).
- NewsLarge: a joint version of multiple NewsCrawl dumps in Spanish released as part of the 2019 Conference on Machine Translation (WMT) (Barrault et al., 2019).
- NewsSmall: in order to make the comparison fairer, we trained a language model with a subset of NewsCrawl that had the same amount of tokens as MedlinePlus.

Table 7 shows the different sizes of the corpora used to build the three language models. These language models were trained with KenLM and a window size of 5 tokens. Apart from that, the parameters documented by Chollampatt and Ng (2018) were set to train the GEC models.

Finally, we trained MLConv with different sets of training data. Given the small size of IMEC, we experimented with oversampling and the incorporation of TMAE to the training data. During the early experimentation phase, IMEC's training section was oversampled with orders of magnitude from 5 to 100, as reported later in the discussion (§4.3). In this work, we only show the 4 best performing combinations of datasets, described in Table 8. From this point on, we will refer to the different oversampling points as IMEC$_{\times N}$, N being the number of repetitions.

Corpus	# Lines	# Tokens
MedlinePlus	445,140	6,461,483
NewsSmall	220,000	6,501,721
NewsLarge	51,833,058	1,588,491,570

Table 7: Size of the corpora used to create the language models.

[5]https://dumps.wikimedia.org/

Corpus train	# Lines	# Tokens
IMEC	7,506	122,812
IMEC$_{\times 75}$	562,950	9,210,900
IMEC + TMAE	2,333,350	51,833,425
IMEC$_{\times 15}$ + TMAE	2,438,434	3,552,793

Table 8: Size of the different training sets for MLConv; the number of tokens correspond to the incorrect partitions.

4.2 Results

Table 9 shows the overall results of the experiments. For each set of experiments, the table indicates the system evaluated and the training data and configuration or re-ranking model used.

On its own, the spellchecker baseline achieves acceptable but low results. Adding specialized, in-domain vocabulary obtained from the training section of IMEC greatly boosts performance. However, any attempts at re-ranking the results, either using the language model trained on the MedlinePlus corpus or capping the suggestions at a given Levenshtein distance, seem to only interfere with Aspell's own ranking and lowers performance. In general, using Aspell returns decent precision scores but really low recall.

The results of training the MLConv network using IMEC are overall better than those obtained by the baseline, even if the corpus' size is small. This improvement is especially appreciated in terms of recall. For each experiment, the effects of re-ranking the output sentences are shown. It seems to have a positive effect, giving a performance boost in comparison to the raw output.

TMAE, which relies on data augmentation as explained in §3.2, is also a great asset. When merged with IMEC, it gave a similar boost to precision as just oversampling IMEC (IMEC$_{\times 75}$), although it seems that recall suffers a little in comparison.

In general, the best performing system is achieved when combining IMEC$_{\times 15}$ with TMAE and re-ranking with MedlinePlus.

4.3 Discussion

The overall picture of the experiments is that neural networks work much better for GEC than spellcheckers. This is something we expected, as the phenomena contained in the corpus are much wider than spelling errors.

Even then, there are some interesting re-

System	Training data / Configuration	Precision	Recall	$F_{0.5}$
Aspell	*as is*	26.44	16.44	23.57
	+ MedlinePlus	17.27	10.69	15.38
	+ train vocab	52.62	14.99	35.03
	+ MedlinePlus + train vocab	30.01	08.59	20.06
	Lev=1 + MedlinePlus	37.95	14.65	28.79
	Lev=1 + MedlinePlus + train vocab	54.80	13.23	33.65
MLConv	IMEC	42.36	41.26	42.14
	IMEC + MedlinePlus	45.23	38.79	43.78
	IMEC + NewsSmall	45.21	38.27	43.63
	IMEC + NewsLarge	46.41	41.03	45.22
	$IMEC_{\times 75}$	53.63	45.81	51.86
	$IMEC_{\times 75}$ + MedlinePlus	62.09	42.71	56.92
	$IMEC_{\times 75}$ + NewsSmall	62.01	42.94	56.95
	$IMEC_{\times 75}$ + NewsLarge	59.85	45.52	56.31
	IMEC + TMAE	62.57	35.43	54.26
	IMEC + TMAE + MedlinePlus	62.14	37.11	54.75
	IMEC + TMAE + NewsSmall	64.00	36.14	55.45
	IMEC + TMAE + NewsLarge	63.89	38.42	56.41
	$IMEC_{\times 15}$	73.71	**59.19**	70.26
	$IMEC_{\times 15}$ + TMAE + MedlinePlus	**76.00**	55.87	**70.89**
	$IMEC_{\times 15}$ + TMAE + NewsSmall	75.30	55.94	70.43
	$IMEC_{\times 15}$ + TMAE + NewsLarge	75.81	55.64	70.69

Table 9: Results of GEC in the IMEC test split. The best results of each experiment set are marked with an underline; the best results overall are highlighted in boldface.

marks that could be made about the baseline. Firstly, Aspell's own ranking system is solid enough that attempting to add any extra layer of suggestion classification only hinders it. Secondly, using in-domain vocabulary almost doubles precision. This highlights the importance of including in-domain vocabulary when dealing with such specialized texts.

An important aspect of our experiments was the re-ranking of the neural networks' output. Unfortunately, each system achieved its best results using a different language model, therefore a conclusion cannot be drawn as to exactly how much the language model's training data's size and domain matter. It is clear, though, that re-ranking is a valuable step, as it improves the system's performance in every single experiment. An argument could be made that, in some cases, precision benefits more than recall from this process. This is not inherently bad, given that in GEC it is preferable to offer good corrections than to suggest dubious corrections for every mistake.

Regarding the creation of the TMAE corpus, it could be said that it was created in a probabilistic way. Interestingly enough, the results of the model that uses it are coherent with the theory presented by Felice (2016) that states that probabilistic generation of synthetic errors increases precision while decreasing recall.

We would also like to provide some insight into the oversampling process. The training section of the IMEC corpus was repeated a different number of times (multiples of 5 up to a 100). A new model was trained with each of them to explore how performance changed. The experiments showed immediate improvement, with a steady increase as the number of repetitions increase and a peak at 75 repetitions. However, after that, the model's performance greatly decreases.

For the joint oversampled IMEC + TMAE model, however, IMEC was repeated only 15 times. After a few experiments, it seemed apparent that a higher number of repetitions seemed to not work as well when combined with more data.

ERRANT's evaluation system also allows us to look at each system's performance individually, as it returns precision, recall and $F_{0.5}$ for each error type. Due to space issues, we are not able to include full tables for each system. These are some highlights:

- As expected, our best system ($IMEC_{\times 15}$ + TMAE) has the best performance in most categories.
- All neural systems return better results for the SPELL category than the spellchecker used for the baseline, usually over 65.00 $F_{0.5}$ score. The ORTH category generally returns really good results across all systems too.
- No system is able to correct ADJ:INFL errors at all (there were few examples in the corpus to begin with), and only $IMEC_{\times 15}$ + TMAE is able to correct some DET:INFL errors. This behaviour is also shown in the VERB:FORM and VERB:SVA categories. This suggests that our convolutional neural network is not able to learn how agreement works and that it may be better suited for languages that are less morphologically rich than Spanish.
- Some of the less frequent categories, such as word order, which has only 4 instances, are not learnt at all by any of the systems.

Finally, an interesting fact that can be appreciated upon manual error analysis is that sometimes the models return correct examples that are not evaluated as such, since they differ from the gold standard. For instance, the system may insert the verb 'presentar' (*to present*) instead of 'mostrar' (*to show*) when a verb is missing. This is actually correct but, due to the lack of multiple annotations for each sentence, it is evaluated as incorrect. Extending our corpus with more data and multiple annotators is left as future work.

5 Conclusion

In conclusion, this paper presents a first approach to Grammatical Error Correction for health records in Spanish. This is a topic that has not been previously explored, but that we consider may have a great impact. Health records are the main form of communication between health specialists and patients, but their form is a flawed aspect that usually contains multiple orthographic and grammatical problems.

GEC may be a helpful solution to this problem. This work introduces the IMEC (*Informes Médicos en Español Corregidos*) corpus–which is made up of over 10,000 manually corrected sentences from health records–as well as the TMAE (*Textos Médicos Aumentados con Errores*) corpus, a parallel collection of over 2 million sentences from the clinical domain augmented with errors.

Additionally, we present extensive experimentation with a Multilayer Convolutional Encoder-Decoder (Chollampatt and Ng, 2018) and the corpora generated. The results show promising results in this line and suggest that it is possible to obtain good results even with small datasets.

As future work, one of the most important steps we would like to take is to expand the IMEC corpus, not only with more data but also with more annotators. Given the subjectivity this field shows, having multiple possibilities for a correction is almost compulsory to avoid false negatives like the ones described at the end of §4.3.

Another significant step would be to test the impact this type of correction has on other NLP tasks via extrinsic evaluation on information extraction or anonymization systems. If clinical GEC systems make text less noisy, it may prove helpful for text processing in general.

Finally, we plan on performing new experiments with more competitive systems based on the Transformers architecture and large pre-trained language models, which have achieved a widespread success in virtually every NLP task.

Acknowledgments

This work has been supported by Vicomtech and partially funded by the projects DeepText (KK-2020-00088, SPRI, Basque Government) and DeepReading (RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE). We also want to thank Olatz Pérez de Viñaspre, who has collaborated in the research behind this article and whose contributions have been essential.

References

Aguilar Ruíz, M. J. 2013. Las normas ortográficas y ortotipográficas de la nueva Ortografía de la lengua española (2010) aplicadas a las publicaciones biomédicas en español: una visión de conjunto. *Panace@: Revista de Medicina, Lenguaje y Traducción*, XIV(37):101–120.

Atkinson, K. 2020. GNU Aspell 0.61 documentation.

Barrault, L., O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Bello Gutiérrez, P. 2016. Aprendiendo a redactar mejor tus informes. *Curso de Actualización Pediatría*, pages 391–400.

Beloki, Z., X. Saralegi, K. Ceberio, and A. Corral. 2020. Grammatical error correction for basque through a seq2seq neural architecture and synthetic examples. *Procesamiento del Lenguaje Natural*, 65:13–20, September.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boletín Oficial del Estado. 2015. Real decreto 9/2015, de 6 de febrero, por el que se regula el registro de actividad de atención sanitaria especializada.

Bryant, C. 2019. *Automatic annotation of error types for grammatical error correction.* University of Cambridge.

Bryant, C., M. Felice, Ø. E. Andersen, and T. Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Chollampatt, S. and H. T. Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5755–5762, New Orleans, Louisiana, USA. AAAI Press.

Dahlmeier, D., H. T. Ng, and S. M. Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Davidson, S., A. Yamada, P. Fernandez Mira, A. Carando, C. H. Sanchez Gutierrez, and K. Sagae. 2020. Developing NLP tools with a new corpus of learner Spanish. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 7238–7243, Marseille, France. European Language Resources Association.

Felice, M. 2016. Artificial error generation for translation-based grammatical error correction. Number 895.

Felice, M., C. Bryant, and T. Briscoe. 2016. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, Osaka, Japan. The COLING 2016 Organizing Committee.

Gamon, M., J. Gao, C. Brockett, A. Klementiev, W. B. Dolan, D. Belenko, and L. Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, pages 449–456. Asia Federation of Natural Language Processing.

Granger, S. 1998. *The computer learner corpus: a versatile new source of data for SLA research.* na.

Grundkiewicz, R., M. Junczys-Dowmunt, and K. Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 252–263, Florence, Italy. Association for Computational Linguistics.

Heafield, K. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Intxaurrondo, A. 2018. SPACCC. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Lee, J. and S. Seneff. 2008. Correcting misuse of verb forms. In *Proceedings of ACL-08: HLT*, pages 174–182, Columbus, Ohio, USA. Association for Computational Linguistics.

Lima López, S., N. Pérez, M. Cuadros, and G. Rigau. 2020. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC2020)*, pages 5772–5781, Marseille, France. European Language Resources Association.

Macdonald, N. H. 1983. Human factors and behavioral science: The UNIX Writer's Workbench software: Rationale and design. *The Bell System Technical Journal*, 62(6):1891–1908.

Ministerio de Sanidad. 2018. Recursos físicos, actividad y calidad de los servicios sanitarios.

Mizumoto, T., Y. Hayashibe, M. Komachi, M. Nagata, and Y. Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of COLING 2012: Posters*, pages 863–872, Mumbai, India. The COLING 2012 Organizing Committee.

Náplava, J. and M. Straka. 2019. Grammatical error correction in low-resource scenarios. In *Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text*, pages 346–356, Hong Kong, China. Association for Computational Linguistics.

Ng, H. T., S. M. Wu, T. Briscoe, C. Hadiwinoto, R. Susanto, and C. Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14. Association for Computational Linguistics.

Ng, H. T., S. M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault. 2013. The conll-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12. Association for Computational Linguistics.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Omelianchuk, K., V. Atrasevych, A. Chernodub, and O. Skurzhanskyi. 2020. GECToR – Grammatical Error Correction: Tag, Not Rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA. Association for Computational Linguistics.

Rei, M., M. Felice, Z. Yuan, and T. Briscoe. 2017. Artificial error generation with machine translation and syntactic patterns. *CoRR*, abs/1707.05236.

Richardson, S. D. and L. C. Braden-Harder. 1988. The experience of developing a large-scale natural language text processing system: Critique. In *Second Conference on Applied Natural Language Processing*, pages 195–202, Austin, Texas, USA. Association for Computational Linguistics.

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

Shen, L., A. Sarkar, and F. J. Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the*

Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Tajiri, T., M. Komachi, and Y. Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea. Association for Computational Linguistics.

Terroba Reinares, A. R. 2015. *Mejora de la calidad del informe clínico de alta hospitalaria desde el punto de vista lingüístico*. Universidad de La Rioja.

Tetreault, J., J. Foster, and M. Chodorow. 2010. Using parse features for preposition selection and error detection. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 353–358, Uppsala, Sweden. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pages 5998–6008.

Villegas, M., A. Intxaurrondo, A. Gonzalez-Agirre, M. Marimon, and M. Krallinger. 2018. The MeSpEN resource for English-Spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. In *Proceedings of the LREC 2018 Workshop "MultilingualBIO: Multilingual Biomedical Text Processing"*, pages 32–39. European Language Resources Association.

Xie, Z., A. Avati, N. Arivazhagan, D. Jurafsky, and A. Y. Ng. 2016. Neural language correction with character-based attention. *CoRR*, abs/1603.09727.

Yannakoudakis, H., Ø. E. Andersen, A. Geranpayeh, T. Briscoe, and D. Nicholls. 2018. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31(3):251–267.

Yannakoudakis, H., T. Briscoe, and B. Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.